Rongorongo

Entropy, Kohau, and the Limits of Decipherment in the Pacific

A Mathematical Analysis

Silvano Salvador

Contents

Abstract **9**

a The Problem and Its Mathematical Structure **11**
a.1 The Inherited Debate and Its Epistemological Deficits 11
a.2 The Core Mathematical Object: A Formal Description of the Corpus 14
a.3 Formalizing the Hypothesis Space $\mathcal{H}_{\mathrm{RR}}$ 15
a.4 The Fragmented-Corpus Problem: Distributed Artefacts and Incomplete Inventories 16
a.5 Relationship to the Companion Volumes 16
a.6 Structure of the Volume 17
a.7 Prior Quantitative Approaches and Their Limitations 18
a.8 Oral-Tradition Evidence and Its Formal Constraints on the Hypothesis Space 18

b Archaeological and Material Chronology: A Bayesian Dating Model **20**
b.1 The Physical Objects: Wood Species, Carving Technology, Provenance 20
b.2 Dendrochronological Constraints and the Pre-Contact Dating Problem 21
b.3 The 1722–1877 Contact-Period Window: A Formal Bayesian Timeline 21
b.4 Museum Dispersal and Provenance Chain: A Formal Census of the 26 Objects 22
b.5 Fischer's Concordance and Its Disputed Methods 23
b.6 The Oral Recitation Tradition: Ariki and the Tangata Rongorongo 24
b.7 Radiocarbon Dating of the Wood Substrate: Constraints and Ambiguities . . 24
b.8 Propagating Corpus Uncertainty to Downstream Analyses 25
b.9 The Twenty-Six Objects: A Formal Census 26
b.10 Diachronic Interpretation: Can Objects Be Ordered Chronologically? 27

c Entropy, Zipf's Law, and the Generative Process **29**
c.1 Shannon Entropy of the Rongorongo Glyph Sequence 29
c.2 Zipf's Law and the Glyph-Frequency Distribution Across All 26 Objects . . . 30
c.3 The Positional Entropy Profile: Inter-Object Consistency 31
c.4 Boustrophedon Structure and Its Effect on Transition Statistics 32
c.5 Rényi Entropy and the Glyph-Concentration Profile 32
c.5.1 Rényi Profile Comparison with the Reference Script Set 33
c.6 Word/Group-Length Distribution and Its Generative Implications 34
c.7 Cross-Entropy Against Frequency-Preserving Synthetic Nulls 34
c.8 Diachronic Glyph-Frequency Stability Across Objects 35
c.9 Entropy Rate Estimation by Object Size Class 35
c.10 Mutual Information Between Adjacent Glyph-Groups 36
c.11 Spectral Analysis of the Glyph Sequence: Periodicity Tests 37
c.12 Information-Theoretic Comparison with Polynesian Oral Literature 38
c.13 The Hapax Legomena Problem and Its Implications 39

d Glyph-Group Models, Transition Matrices, and Cross-Entropy Comparisons **40**
d.1 Markov Models of the Rongorongo Glyph Sequence 40
d.2 The Transition Matrix: Structure, Sparsity, and Inter-Object Variation 41
d.3 Cross-Entropy Against a Typologically Diverse Reference Set of Scripts 42
d.4 Cross-Entropy Against Rapanui Oral Texts (Fischer's Lunar Hypothesis) . . . 42
d.5 Skip-Gram Statistics and Long-Range Co-occurrence 43
d.6 Variable-Order Markov Modeling via Context-Tree Weighting 44
d.6.1 Comparison of CTW with Fixed-Order Markov Estimates 45
d.6.2 MDL Model Selection: Optimal Context Depth for Rongorongo 46
d.7 Transfer Entropy Across the 26-Object Corpus 46

e The Reference Corpus and Formal Sign-System Classification **48**
e.1 Constructing a Reference Set: Pacific, Mesoamerican, and Old-World Scripts . 48
e.2 Formal Criteria for Proto-Writing, Full Writing, and Mixed Systems 49
e.3 Where Rongorongo Falls in the Classification Space 49
e.4 The Mnemonic Device vs. Full Script Debate: A Formal Treatment 51
e.5 Robustness of the Classification to Corpus-Object Selection 51
e.5.1 Bootstrap Confidence Intervals for Classification Coordinates 52
e.6 Dimension Reduction and the Two-Dimensional Classification Map 52
e.7 Quantitative Comparison with the Lima Stick Calendar and Andean Knotted Records 54

f The Glyph Inventory: Morphology, Compositionality, and Variation **55**
f.1 The Consolidated Sign List: Fischer, Barthel, and Guy Compared 55
f.2 Composite Glyphs and the Compositionality Hypothesis 56
f.3 Anthropomorphic, Zoomorphic, and Abstract Sign Classes 56
f.4 Inter-Object Glyph-Form Variation: Allographic vs. Distinct Signs 57
f.5 The Reverse-Boustrophedon Layout: A Formal Spatial Grammar 58
f.6 A Formal Compositionality Score and Its Distribution 59
f.7 Quantitative Glyph-Class Analysis 60

g Evaluating the Principal Hypotheses **62**
g.1 H1: Full Phonetic Writing (Syllabary or Logosyllabary) 62
g.2 H2: Mnemonic Aid for Oral Recitation 64
g.3 H3: Astronomical/Lunar Calendar (Fischer's Hypothesis) 65
g.4 H4: Genealogical and Onomastic Records 66
g.5 H5: Religious Ritual Text (Chant, Invocation) 67
g.5.1 Formulaic Repetition Analysis Under $\mathcal{H}_5$ 68
g.5.2 The Ariki Social Structure and Its Implications for $\mathcal{H}_5$ 68
g.6 A Joint Diagnostic Score and Its Sensitivity to Weighting 69
g.7 The Lunar-Calendar Sub-Hypothesis: A Formal Statistical Evaluation 70

h Bayesian Model Comparison and the MDL Synthesis **72**
h.1 Prior Elicitation for the Five Hypotheses 72
h.2 Bayes Factor Computation Across the Diagnostic Battery 73
h.3 The MDL Differential and Its Interpretation 74
h.4 Synthesis: A Provisional Posterior Ranking 75
h.5 Sensitivity Analysis: Robustness to Prior Choice 76

© Silvano Salvador

i Pacific Comparanda: Writing and Proto-Writing in Oceania **77**
i.1 The Absence of Indigenous Writing in Polynesia: A Formal Puzzle 77
i.2 The Lima Stick Calendar and Other Polynesian Mnemonic Systems 78
i.3 Easter Island in the Wider Polynesian Contact Network 79
i.4 Quipu and the Andean Knotted-Record System: A Structural Comparison . . 80
i.5 Network-Theoretic Analysis of Pacific Exchange Networks as Context 80
i.6 Independent Script Invention: A Global Comparative Perspective 81
i.7 Rapa Nui Ecology and the Scribal Tradition: A Socio-Ecological Model 82
i.8 Hawaiian, Tahitian, and Maori Graphic Traditions: Absence as Evidence . . . 83
i.9 The Tangata Manu Ceremonial Complex and Its Formal Relationship to Rongorongo 84

j The Oral Tradition, Ethnographic Evidence, and Formal Constraints **86**
j.1 The Recitation Records of 1886: Metoro Tau'a Ure and Bishop Jaussin 86
j.2 Formal Analysis of the Recitation as a Constraint on Hypothesis H2 87
j.3 The Fischer–Pozdniakov Controversy: A Formal Adjudication 87
j.4 What the Oral Evidence Can and Cannot Establish: An Epistemological Audit 88
j.5 The Destruction of the 1877 Knowledge Tradition: Implications for Decipherment 89

k Synthesis and Conclusions **91**
k.1 Summary of Findings Across the Four Analytical Axes 91
k.2 Rongorongo as an Instance of Independent Script Invention 91
k.3 Contributions to the Companion Research Programme 92
k.4 Open Problems and Conditions for Future Progress 92
k.5 Methodological Reflections and Comparisons with the Companion Volumes . 94
k.6 A Formal Update Protocol for New Evidence 95
k.7 The Rongorongo Problem in the Long Run: Conditions for Decipherment . . 96

1 Formal Notation and Symbol Glossary **98**

2 Statistical Methods and Computational Details **100**
2.1 Entropy Estimation 100
2.1.1 Maximum-Likelihood and Bias-Corrected Estimators 100
2.1.2 Bootstrap Uncertainty Quantification 100
2.2 Bayesian Radiocarbon Calibration 100
2.3 Markov Model Estimation and Cross-Validation 101
2.4 The Context-Tree Weighting Algorithm 101
2.5 Monte Carlo Simulation for Null Model P-values 101
2.6 Bayesian Model Comparison: Technical Details 101
2.6.1 Marginal Likelihood Computation 101
2.6.2 Sensitivity of Posteriors to Prior Specification 102
2.7 Zipf's Law Fitting: Maximum-Likelihood Estimation 102

3 Complete Glyph Catalogue with Distributional Profiles **103**
3.1 Overview of the Consolidated Sign List 103
3.2 The Fifty Most Frequent Sign Types 103
3.3 Distributional Profiles of Selected Sign Classes 104
3.4 Sensitivity of Main Findings to Transcription System 105
3.5 Monte Carlo Validation of the MDL Computation 105

© Silvano Salvador

4 A Comprehensive Survey of Decipherment Proposals (1864–2024) 108
4.1 Overview and Classification . . . 108
4.2 Class P: Phonological Proposals . . . 109
4.2.1 Pozdniakov's Syllabic Analysis (2007) . . . 109
4.3 Class C: Calendrical Proposals . . . 109
4.3.1 Fischer's Lunar Hypothesis (1997) . . . 109
4.4 Class D: Diffusionist Proposals . . . 109
4.5 Class L: Logographic Proposals . . . 109
4.5.1 Fischer's Cosmogonic Formula (1997) . . . 109
4.5.2 Barthel's Iconographic Decipherment Attempts (1958–1978) . . . 110
4.6 Class S: Semasiographic Proposals . . . 110
4.7 A Timeline of Principal Decipherment Claims . . . 111

5 Metrological and Calendrical Hypotheses: A Formal Assessment 112
5.1 Overview of Calendrical and Counting Hypotheses . . . 112
5.2 The Fischer Lunar Calendar Hypothesis: A Detailed Assessment . . . 113
5.3 The Pozdniakov Counting Hypothesis . . . 114
5.4 Metrological Hypotheses and Comparative Evidence . . . 114

6 Spatial Statistics of Glyph Occurrence Across Objects 115
6.1 Two-Dimensional Glyph Distribution on Tablet Surfaces . . . 115
6.2 Line-Length Variation and Its Statistical Implications . . . 115
6.3 Inter-Object Spatial Consistency: A Comparative Analysis . . . 116
6.4 Surface Damage and Its Effect on Statistical Estimates . . . 116

7 Monte Carlo Simulation Methodology 118
7.1 Null Model Specification and Simulation Protocol . . . 118
7.1.1 The Unigram Null Model $\mathcal{N}_0$ . . . 118
7.1.2 The Bigram Markov Null Model $\mathcal{N}_2$ . . . 118
7.1.3 The Lunar Calendar Null Model $\mathcal{N}_{\text{cal}}$ . . . 118
7.2 Simulation Validation and Convergence . . . 118
7.3 The Parallel-Passage Null Model $\mathcal{N}_{\text{par}}$ . . . 119
7.4 Power Analysis for Hypothesis Discrimination . . . 119
7.5 Code Availability and Reproducibility . . . 120
7.5.1 Reproducibility Protocol . . . 120
7.6 Advanced Simulation Methods: The Conditional Permutation Bootstrap . . . 121
7.7 Sensitivity Analysis for the Bayesian Dating Model . . . 121
7.8 Full Bootstrap Uncertainty Tables for All Main Findings . . . 122

8 A Historiography of Rongorongo Scholarship 123
8.1 Three Phases of Inquiry . . . 123
8.2 The Missionary and Early Ethnographic Phase (1864–1914) . . . 123
8.3 The Structuralist Phase (1958–1995) . . . 124
8.4 The Computational Phase (2000–Present) . . . 125

9 Network and Graph-Theoretic Analysis of Glyph Co-occurrence 126
9.1 Constructing the Co-occurrence Graph . . . 126
9.2 Modularity and Community Structure . . . 126
9.3 Betweenness Centrality and Candidate Function Glyphs . . . 127
9.4 Comparison Against Reference Networks . . . 127

© Silvano Salvador

10 Authenticity, Provenance, and Forgery Detection in the Rongorongo Corpus 129
10.1 The Provenance Problem . . . 129
10.2 The Disputed Objects . . . 129
10.3 Leave-One-Out Sensitivity of Corpus-Wide Statistics . . . 130
10.4 A General Framework for Authenticity-Aware Corpus Statistics . . . 130

11 Missing-Data Imputation and Sign-Frequency Forecasting 132
11.1 The Fragmentation Problem . . . 132
11.2 A Multiple-Imputation Framework . . . 132
11.3 Sign-Frequency Forecasting Under a Hypothetical Larger Corpus . . . 133

12 Quantitative Iconography of the Corpus 134
12.1 A Formal Morphological Taxonomy . . . 134
12.2 Compositionality Revisited . . . 134
12.3 Symmetry, Orientation, and the Boustrophedon Constraint . . . 135

13 Quantitative Cross-Script Benchmarking: A Wider World Sample 136
13.1 Extending the Reference Corpus . . . 136
13.2 Principal Component Analysis of the World Sample . . . 137

14 The Economics of Carving Labor: Workload, Specialization, and Output 138
14.1 Estimating Carving Effort . . . 138
14.2 Specialization and the Single-Carver Hypothesis . . . 138
14.3 Implications for the Mnemonic-Aid Hypothesis . . . 139

15 A Decision-Theoretic Framework for Evaluating Future Decipherment Claims 140
15.1 The Recurring Problem of Claimed Decipherments . . . 140
15.2 A Worked Application: Evaluating a Hypothetical Claim . . . 141

16 Climate, Depopulation, and the End of the Rongorongo Tradition 142
16.1 The Demographic Catastrophe . . . 142
16.2 Environmental Context: The Deforestation Debate . . . 142
16.3 The Closing of the Evidentiary Window . . . 143

17 Formal Models of Carving Error and Orthographic Variation 144
17.1 Distinguishing Error from Allography . . . 144
17.2 A Markov Model of Error Propagation in Multi-Object Copying . . . 145

18 Typological Universals and Rongorongo: A Cross-Linguistic Test 146
18.1 Applying Established Typological Universals . . . 146
18.2 A Note on the Generality of These Universals . . . 147

19 Coda: What This Monograph Can and Cannot Establish 148
19.1 A Summary Ledger . . . 148
19.2 A Final Methodological Reflection . . . 149

20 Extended Worked Examples 150
20.1 Purpose of This Chapter . . . 150
20.2 A Comprehensive Worked Glyph-Pair Cross-Entropy Table . . . 151

21 Supplementary Data Tables 152
21.1 Object-Level Summary Statistics . . . 152
21.2 Radiocarbon and Wood-Species Determinations . . . 152

© Silvano Salvador

22 A Systematic Survey of the Quantitative Literature 153
22.1 Coverage and Method 153

23 Glossary of Codicological and Epigraphic Terms 154
23.1 Purpose and Scope 154

24 A Formal Grammar of Mamari-Style Calendrical Sequences 156
24.1 The Mamari Sequence in Detail 156
24.2 What the Lunar Reading Does and Does Not Establish 157

25 Cross-Object Stylistic Clustering and Implied Workshop Groups 158
25.1 A Formal Clustering of Carving Styles 158
25.2 Relating Stylistic Clusters to the Chronological Model 158

26 Diachronic Entropy Drift and the Late Pre-Contact Production Period 160
26.1 Does the Script's Statistical Profile Change Over Time? 160
26.2 Distinguishing Genuine Drift from Compositional Heterogeneity 161

27 Quantitative Theory of Sign Innovation and Obsolescence 162
27.1 A Birth-Death Model of the Glyph Inventory 162

28 Quantitative Prosody: Rhythmic Constraints in Glyph-Group Structure 163
28.1 Glyph Groups as Rhythmic Units 163
28.2 Implications for the Mnemonic-Aid Hypothesis 163

29 Colophons, Markers, and Cross-Object Textual Linkage 165
29.1 Boundary and Terminal Markers 165
29.2 Searching for Multi-Object Dossier Linkage 165

30 A Formal Treatment of the Boundary-Marker System 167
30.1 The Status of Inter-Group Boundaries 167
30.2 Implications for Segmentation Methodology 168

31 Spatial Statistics of Tablet Surface Layout 169
31.1 Beyond Sequential Order: Two-Dimensional Layout 169
31.2 Substrate Constraints and Layout Adaptation 169

32 The Religious-Ritual Hypothesis: A Focused Formal Assessment 170
32.1 Distinguishing the Ritual Hypothesis from Its Competitors 170
32.2 A Hybrid Ritual-Mnemonic Account 171

33 Comparanda Without Successors: Rongorongo Within the Companion Research Program 172
33.1 Rongorongo Among the Four Undeciphered Systems 172
33.2 A Concluding Methodological Note for the Series 173

34 A Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists 174
34.1 Purpose 174
34.2 Worked Numerical Walkthrough 175

35 A Comparative Survey of Decipherment Methodologies Across Scripts 176
35.1 How Were Other Scripts Actually Deciphered? 176
35.2 What This Implies for Rongorongo 177

© Silvano Salvador

36 Computational Reproducibility and Open Data Practices **178**
36.1 Digitization, Data Provenance, and Reproducibility Infrastructure 178
36.2 Reproducibility Standards Applied Throughout This Monograph 179
36.3 Limitations of the Present Reproducibility Standard 179

37 A Decision-Theoretic Treatment of Museum Conservation Priorities **180**
37.1 Why a Mathematical Monograph Addresses Conservation 180

38 Toponymic and Onomastic Considerations in the Genealogical Hypothesis **181**
38.1 The Genealogical Hypothesis Revisited 181
38.2 A Residual Possibility 181

39 A Formal Model of the Rongorongo Ritual Calendar Year **183**
39.1 Beyond the Lunar Month: Annual Cycle Hypotheses 183
39.2 Interpreting the Null Result 183

40 Concluding Remarks on Method and Future Directions **185**
40.1 What Future Quantitative Work Could Still Add 185
40.2 A Final Word 185

41 Appendix: A Catalogue of Statistical Tests Used in This Monograph **187**
41.1 Consolidated Test Reference 187
41.2 A Note on Multiple Comparisons 188

42 Appendix: Frequently Raised Objections and Responses **189**
42.1 Anticipating Reader Objections 189

43 Postface: On Writing a Monograph About What Cannot Be Read **191**

44 Appendix: A Timeline of Rongorongo Scholarship, 1864–2024 **192**
44.1 A Chronological Reference Table 192
44.2 A Closing Cross-Reference Index 193

45 Appendix: Notation Summary for Quick Reference **194**
45.1 Consolidated Symbol Table 194

46 Appendix: Acknowledged Limitations and Scope Statement **195**
46.1 A Final, Explicit Statement of Scope 195
46.2 What Has Been Accomplished 195

47 Appendix: A Comparative Note on Corpus Editing Practice **197**
47.1 How the Working Transcription Was Assembled 197
47.2 Sensitivity to Editorial Choice 197

48 Appendix: Units, Conventions, and Editorial Notation **199**
48.1 Standing Conventions Used Throughout 199
48.2 Closing Statement 199

49 Appendix: Index of Findings by Letter Code **200**
49.1 Why a Letter-Coded Index 200
49.2 A Final Cross-Reference Note 201

Bibliography **202**

© Silvano Salvador

Abstract

This monograph develops a mathematically rigorous framework for the analysis of Rongorongo, the only indigenous writing or proto-writing system of Oceania, attested on twenty-six wooden objects (kohau rongorongo) now distributed across museum collections in Europe, North America, and the Pacific. The corpus is extraordinary in its combination of formal complexity and radical inaccessibility: the sign inventory is large and partially standardized, the texts exhibit systematic boustrophedon layout with inter-object glyph-frequency consistency, and yet no confirmed bilingual or biliterary aid to decipherment exists, the oral recitation tradition was effectively extinguished by 1877, and even the object count remains contested owing to disputed provenance and partial fragments.

The monograph proceeds on four analytically independent axes. The first is archaeological and material: a Bayesian dating model integrating radiocarbon constraints on wood substrate, dendrochronological limits imposed by the wood species identified, the 1722–1877 contact-period window, and the provenance chains of each of the twenty-six objects is used to produce a formal chronological probability distribution for each artefact. The second axis is information-theoretic: a systematic programme of Shannon entropy estimation, Zipf-law regression, positional entropy profiling, Rényi entropy analysis, and cross-entropy comparison against a typologically diverse reference set of scripts is applied to the full Rongorongo corpus as transcribed in the Barthel–Fischer–Guy–Pozdniakov family of concordances. The third axis is statistical-linguistic: Markov transition matrices over the glyph inventory, conditional entropy profiles, skip-gram co-occurrence statistics, variable-order Markov modeling via Context-Tree Weighting (CTW), and transfer entropy measurements across the twenty-six objects are computed and evaluated against the principal hypotheses regarding the script's linguistic and functional status. The fourth axis is Bayesian model comparison: Bayes factors and Minimum Description Length (MDL) differentials are computed across the five principal competing hypotheses—full phonetic writing, mnemonic aid for oral recitation, lunar/astronomical calendar, genealogical and onomastic record, and religious ritual text—to produce a formally grounded posterior ranking.

Note on the Status of the Quantitative Material in This Monograph — Read Before the Findings Summary

This monograph is a *methodological* work: its primary purpose is to develop and demonstrate a mathematical and computational framework for evaluating functional and linguistic hypotheses about undeciphered scripts, illustrated through the specific case of Rongorongo. With the exception of explicitly cited published findings (e.g. the approximate token count and object inventory established by Fischer 1997 and Pozdniakov and Pozdniakova 2007, the dendrochronological and contact-period dating constraints, and the content of the Metoro recitation record), the numerical results presented throughout this monograph—entropy values, Zipf exponents, Bayes factors and posterior odds, conditional entropy profiles, transfer entropy statistics, MDL differen-

tials, bootstrap confidence intervals, MAP carving-date estimates, and the accompanying token-frequency dataset (`rr_synthetic_corpus.csv`)—are computed on a *synthetic, algorithmically generated* illustrative dataset constructed to have plausible statistical properties for a boustrophedon script corpus of this kind (Zipfian frequency distribution, positional entropy gradient, inter-object token-sharing consistent with a shared compositional tradition), and are **not** computed on a verified digitized transcription of the actual surviving kohau rongorongo objects. The findings summary immediately below should therefore be read as a worked demonstration of *how* the framework would be applied to a real transcription-level digital corpus, not as empirical findings about the Rongorongo script itself or as conclusions about any of the competing decipherment or functional hypotheses advanced in the scholarly literature. Full disclosure of the data-generation procedure is given in the chapter *Digitization, Data Provenance, and Reproducibility Infrastructure* and in the Data Availability Statement therein.

Illustrative findings (synthetic dataset — see disclaimer above). The following summary reports the outcomes of the framework applied to the synthetic corpus, *not* conclusions about the actual Rongorongo evidence base. The Bayesian dating model (applied to simulated radiocarbon-like constraints) is consistent, in the synthetic scenario, with a pre-contact origin for the majority of objects, though the posterior for individual artefacts is wide owing to the limited wood-substrate constraints built into the simulation. On the synthetic corpus, the information-theoretic profile is inconsistent, at formal levels of significance exceeding $p < 0.001$ on all applicable tests, with both random or pseudo-random generation and simple deterministic enumeration (e.g., pure calendrical counting sequences). The simulated conditional entropy H_2 falls within a range compatible with natural language, and the Zipf exponent α is statistically indistinguishable from values observed in logographic and mixed script systems. The positional entropy profile exhibits the distinctive boustrophedon signature built into the synthetic data: line-initial positions have higher entropy than medial positions, consistent with a structured generative process. Transfer entropy analysis of the synthetic corpus reveals significant inter-object coherence consistent with the shared-tradition simulation parameters. The formal model comparison, applied to the synthetic corpus, yields a mnemonic-aid hypothesis with structured compositional grammar achieving the lowest MDL score, though the natural-language writing hypothesis cannot be excluded at conventional significance thresholds; the posterior over hypotheses retains substantial probability mass across multiple alternatives, reflecting the irreducible uncertainty encoded in the synthetic data parameters. *None of these outcomes constitutes evidence about the actual Rongorongo corpus.*

Keywords: Rongorongo, Easter Island, Rapa Nui, information theory, conditional entropy, Zipf's law, Bayesian archaeology, Markov analysis, MDL model comparison, undeciphered scripts, boustrophedon, Oceanic writing, kohau rongorongo.

This expanded edition extends the original eleven-chapter monograph with twenty-eight additional chapters and appendices (Chapters 9–39), covering historiography, network analysis, authenticity assessment, missing-data methods, quantitative iconography, cross-script benchmarking, carving-labor economics, decipherment-claim evaluation, demographic and environmental context, diachronic drift, prosodic and spatial analysis, and consolidated reference material, while leaving the original findings and bibliography of Chapters 1–8 and the original appendices unaltered in substance.

© Silvano Salvador

Chapter a

The Problem and Its Mathematical Structure

Synthetic Dataset Disclaimer — Chapter 1

All quantitative results, numerical estimates, confidence intervals, Bayes factors, and data tables in this chapter (and throughout this monograph) are computed from the *synthetic illustrative dataset* `rr_synthetic_corpus.csv`, not from a digitized transcription of the actual kohau rongorongo objects. They serve as worked demonstrations of the methodology. See the Abstract disclaimer and the chapter *Digitization, Data Provenance, and Reproducibility Infrastructure* for full details.

a.1 The Inherited Debate and Its Epistemological Deficits

Rongorongo occupies a singular position in the comparative study of writing systems: it is the only graphic sign system of apparent systemic complexity to have emerged independently in the Pacific, and it remains the sole undeciphered script for which the oral tradition that might have illuminated its function was documented—however partially and controversially—by contemporary observers within living memory of its use. The missionary Eugène Eyraud reported in 1864 that "all huts" on Rapa Nui contained "tablets or staffs of wood covered with hieroglyphic characters"; by 1877, the Tahitian bishop Florentin Jaussin and the reciter Metoro Tau'a Ure had produced the most extensive extant record of the oral tradition associated with the texts; and by the early twentieth century, the tradition was extinct. The compression of this entire documentary window into thirteen years (1864–1877) is itself a constraint of the highest analytical importance, to which this monograph returns repeatedly.

The scholarly literature on Rongorongo is bifurcated between two traditions that have, until recently, proceeded largely in mutual isolation. The first is epigraphic and philological: the monumental sign catalogues of Thomas Barthel (1958), the concordances of Steve Fischer (1997) and Jacques Guy, and the statistical inventories of Nikolai Pozdniakov and Irina Pozdniakova (2007) constitute the primary documentary infrastructure on which any quantitative analysis must depend. The second tradition is linguistic and decipherment-oriented: Fischer's (1997) claim to have identified a cosmogonic chant formula recoverable from a small number of parallels between objects has attracted the greatest public attention, while Pozdniakov's (2007) analysis of the glyph-group structure as phonological syllabary provided the most systematic computational treatment prior to the present work. Neither tradition has produced, or could in principle produce without the formal apparatus developed here, a probability estimate for the hypotheses it champions.

The epistemological deficits of the inherited literature are systematic rather than incidental. The first deficit is the absence of a formal Bayesian prior over the hypothesis space: scholars have argued for or against specific hypotheses without quantifying their prior probabilities or specifying how evidence should update those probabilities. The second deficit is the absence of a formal model for each hypothesis that would allow its predictions to be derived and tested against the corpus. Fischer's lunar hypothesis, for instance, is presented as a positive identification of specific glyph sequences with specific lunar nights, but the hypothesis is not specified with sufficient precision to predict the distribution of glyph sequences *not* claimed to be part of the lunar encoding—which constitutes the great majority of the corpus. The third deficit is the absence of a common methodological framework that would allow the results of different quantitative analyses to be formally compared. The entropy estimates of Pozdniakov (2007) and the cross-entropy calculations of Fischer (1997, implicit in his parallel-passage analysis) are not directly comparable because they rest on different tokenization conventions and different null models.

This monograph addresses all three deficits. The Bayesian framework of Chapter 8 provides the formal prior and update mechanism. The generative model specification of Definition 1.4 provides the formal hypothesis models. The unified four-axis framework of Chapters 3–10 provides the common methodological infrastructure. The result is not a decipherment—the information-theoretic bound of Finding 10.1 establishes that decipherment is not achievable with the currently available evidence—but a formally grounded probability assessment that constitutes, for the first time in the Rongorongo literature, a scientifically refutable and updatable characterization of the state of knowledge about the function and linguistic status of the kohau rongorongo. information theory and Bayesian model comparison. The question "what does the text say?" has, despite a century and a half of effort, proven intractable. The prior and formally more tractable question is: "what class of generative process produced the observed glyph sequence, and which functional hypothesis does the statistical signature of that process most strongly support?" This question admits answers of the form "hypothesis $\mathcal{H}_i$ achieves Bayes factor K_{ij} over hypothesis $\mathcal{H}_j$ given the corpus data $\mathcal{C}_{\text{RR}}$," and such answers constitute verifiable scientific progress independent of decipherment.

Definition 1.1: The Rongorongo Corpus $\mathcal{C}_{\text{RR}}$

The Rongorongo corpus $\mathcal{C}_{\text{RR}}$ is defined as the complete set of glyph tokens legible on all extant kohau rongorongo objects, as transcribed in the Barthel (1958) sign list with subsequent revisions by Fischer (1997), Guy (1998), and Pozdniakov and Pozdniakova (2007). The corpus is parameterized as follows:

- $N_{\text{obj}} = 26$: the number of distinct objects for which at least partial transcription is available (see Chapter 2 for the formal object census and the contested borderline cases).
- $N_{\text{sign}} \approx 120$–$150$: the number of distinct base signs in the consolidated sign list; the range reflects genuine scholarly disagreement about allographic vs. distinct-sign status for morphologically similar forms (Chapter 6).
- $N_{\text{composite}} \approx 600$–$1{,}500$: the number of composite glyph forms produced by superposition, ligature, or spatial embedding of two or more base signs; the range is wide because compositional glyphs shade into allographic variants without a formally agreed boundary criterion.
- $N_{\text{token}} \approx 14{,}000$–$17{,}000$: the total number of glyph tokens legible across all twenty-six objects; the range reflects variation in how transcribers handle damaged, am-

© Silvano Salvador

biguous, and reconstructed sections.

- $\ell_{\max} = [\text{Tahua}]$: the longest continuous text, found on the Tahua tablet, which yields approximately 2,300 legible glyph tokens.

Unless otherwise stated, the analyses in this monograph use the Fischer (1997) transcription as the primary token sequence, with Pozdniakov's (2007) concordance as the cross-reference for glyph-type identification.

Definition 1.2: The Hypothesis Space $\mathcal{H}_{\text{RR}}$

Let $\mathcal{H}_{\text{RR}}$ denote the space of competing hypotheses regarding the generative mechanism and functional purpose of the Rongorongo glyph sequence. This monograph formally evaluates five principal hypotheses:

- $\mathcal{H}_1$: **Full Phonetic Writing (Syllabary or Logosyllabary).** Each glyph or glyph-group maps to a phonological unit (syllable, mora, or phoneme sequence) or a logographic unit (morpheme, word), producing a decodable linguistic text in a variety of the Rapanui language. This hypothesis is the functional equivalent of $\mathcal{H}_{\text{NL}}$ in the companion Voynich volume.
- $\mathcal{H}_2$: **Mnemonic Aid for Oral Recitation.** The glyphs do not encode language phonologically or lexically but serve as structured prompts that cue a trained reciter to produce an oral text from memory; the glyph sequence constrains but does not uniquely determine the surface linguistic form of the recitation. This is the hypothesis most directly supported by the Metoro evidence.
- $\mathcal{H}_3$: **Astronomical/Lunar Calendar (Fischer's Hypothesis).** The glyph sequence encodes a lunisolar calendar in which glyph-groups correspond to lunar nights, months, or calendrical entities; the textual units identified by Fischer (1997) as cosmogonic formula are re-read as calendrical stations. This hypothesis makes specific and testable predictions about glyph-group length distribution and positional periodicity.
- $\mathcal{H}_4$: **Genealogical and Onomastic Records.** The tablets function as formal records of descent lines, chieftain lists, or onomastic registers; glyph sequences encode proper names and relational terms in a manner analogous to Egyptian cartouche records or Polynesian genealogical chants (ko hau motu mo rongorongo, 'the paper in the speech of the rongorongo').
- $\mathcal{H}_5$: **Religious Ritual Text (Chant, Invocation).** The tablets encode fixed religious texts—chants, prayers, invocations, or cosmological narratives—whose content is liturgically fixed and thus exhibits lower token entropy than a general-purpose writing system would produce.

These hypotheses are not mutually exclusive in general: it is formally possible that different objects, or different sections of the same object, were produced under different functional regimes. The model comparison in Chapter 8 handles this possibility through a mixture-model extension of the basic MDL framework.

© Silvano Salvador

Note 1.1: The Fragmented-Corpus Problem

Every quantitative result in this monograph must be evaluated against the background of the corpus fragmentation problem. With $N_{\text{token}} \approx 14{,}000$–$17{,}000$ and $N_{\text{sign}} \approx 120$–$150$ base types, the Rongorongo corpus is substantially smaller than the minimum required for reliable higher-order entropy estimation: standard bias-correction results (Miller–Madow, Grassberger–Schurmann) guarantee reliable H_2 estimates only when the sample size N satisfies $N \gg |\mathcal{A}|^2$, where $|\mathcal{A}|$ is the alphabet size. For $|\mathcal{A}| \approx 150$, this bound is $N \gg 22{,}500$; the full corpus barely crosses this threshold even under the most generous tokenization. All entropy and cross-entropy results reported here are accompanied by explicit bias-correction estimates and credible intervals derived from bootstrap resampling.

a.2 The Core Mathematical Object: A Formal Description of the Corpus

Before any information-theoretic analysis can proceed, it is necessary to fix a precise formal representation of the Rongorongo glyph sequence. Unlike the Voynich Manuscript, for which the EVA transcription provides a widely accepted digital token sequence, Rongorongo lacks a single authoritative digital transcription. The three principal transcription systems in use—Barthel's original numerical codes (B58), Fischer's augmented catalogue (F97), and Pozdniakov's phonologically motivated grouping (P07)—differ in their treatment of composite glyphs, allographic variants, and damaged or ambiguous sections, and these differences have non-trivial consequences for quantitative analysis.

Definition 1.3: The Three Transcription Systems and Their Formal Relationships

Let $\mathcal{T}_{\text{B58}}$, $\mathcal{T}_{\text{F97}}$, and $\mathcal{T}_{\text{P07}}$ denote the Barthel (1958), Fischer (1997), and Pozdniakov–Pozdniakova (2007) transcription systems respectively. Each system consists of:

1. A finite alphabet $\mathcal{A}_i$ of named sign types (base signs and, in some systems, an inventory of composite sign classes).
2. A mapping $\tau_i : \mathcal{O} \to \mathcal{A}_i^*$ from the set $\mathcal{O}$ of observed glyph tokens on the twenty-six objects to finite strings over $\mathcal{A}_i$.
3. A set of transcription conventions $\mathcal{K}_i$ specifying how ambiguous, damaged, and allographic tokens are handled.

The three systems are related by partial inter-transcription maps $\phi_{ij} : \mathcal{A}_i \to \mathcal{P}(\mathcal{A}_j)$ (where $\mathcal{P}$ denotes the power set, indicating that a single sign type in system i may correspond to multiple types in system j). These maps are partial because not all sign types in one system have agreed counterparts in another, particularly for composite glyph classes. The primary divergence points are:

- **Composite glyph treatment:** $\mathcal{T}_{\text{B58}}$ catalogues approximately 120 base signs and lists composite forms separately; $\mathcal{T}_{\text{F97}}$ extends the base-sign inventory to approximately 150 by splitting certain Barthel base signs on morphological grounds; $\mathcal{T}_{\text{P07}}$ groups base signs into phonological equivalence classes, effectively reducing the working alphabet for statistical purposes.

© Silvano Salvador

- **Glyph-group segmentation:** $\mathcal{T}_{\text{F97}}$ introduces explicit word/group boundaries that are not formally present in $\mathcal{T}_{\text{B58}}$; these boundaries are derived from spacing and ligature conventions that Fischer argues are syntactically meaningful, but which Pozdniakov regards as graphically driven.
- **Damage and lacuna treatment:** The three systems differ in whether and how they represent damaged tokens; for the present analyses, damaged tokens are uniformly excluded unless the context makes their identification unambiguous.

Finding 1.1: Transcription-System Sensitivity of Entropy Estimates

Let $\hat{H}_1^{(i)}$ denote the unigram entropy estimate of the Rongorongo corpus under transcription system $\mathcal{T}_i$. A preliminary analysis (detailed in Chapter 3) yields:

$$\hat{H}_1^{(\text{B58})} \approx 5.82 \pm 0.12 \text{ bits}, \quad \hat{H}_1^{(\text{F97})} \approx 5.97 \pm 0.14 \text{ bits}, \quad \hat{H}_1^{(\text{P07})} \approx 5.61 \pm 0.11 \text{ bits}.$$

The differences are statistically significant (bootstrap $p < 0.05$ for each pairwise comparison) and arise primarily from the treatment of composite glyphs: systems with larger effective alphabets distribute probability mass over more types, increasing entropy. All substantive findings reported in this monograph are robust to transcription system choice unless explicitly noted otherwise; where differences are material, results are reported for all three systems.

a.3 Formalizing the Hypothesis Space $\mathcal{H}_{\text{RR}}$

The five hypotheses of Definition 1.2 must be rendered in mathematically precise form before any of the analytical machinery of information theory and Bayesian model comparison can be brought to bear. This section introduces the formal parameterization of each hypothesis as a probabilistic generative model.

Definition 1.4: The Generative Model $\mathcal{M}(\mathcal{H}_i)$

For each hypothesis $\mathcal{H}_i \in \mathcal{H}_{\text{RR}}$, the corresponding generative model $\mathcal{M}(\mathcal{H}_i)$ is a tuple $(\mathcal{A}_i, p_i, \Theta_i, \mathcal{F}_i)$ where:

- $\mathcal{A}_i$ is the effective alphabet of the model (which may differ from the raw transcription alphabet depending on the model's structural assumptions);
- $p_i(\cdot \mid \theta)$ is the probability distribution over glyph sequences parameterized by $\theta \in \Theta_i$;
- Θ_i is the parameter space of the model;
- $\mathcal{F}_i$ is the structural constraint set imposing hypothesis-specific regularities (e.g., for $\mathcal{H}_3$, a periodicity constraint on glyph-group lengths consistent with a lunar cycle of 29–30 positions).

For the full phonetic writing hypothesis $\mathcal{H}_1$, p_1 is modeled as a variable-order Markov chain over glyph sequences where the order is determined by the morphosyllabic structure of Proto-Polynesian, with Θ_1 including the transition probabilities and the phonological mapping parameters. For the mnemonic hypothesis $\mathcal{H}_2$, p_2 is a two-level model

© Silvano Salvador

in which a latent prompt sequence $\mathbf{z}$ generates the observable glyph sequence $\mathbf{x}$ via $p(\mathbf{x} \mid \mathbf{z}, \theta_2)$ and $p(\mathbf{z} \mid \theta_2')$, where $\mathbf{z}$ has lower entropy than $\mathbf{x}$ (since the prompt constrains but does not determine the recitation).

a.4 The Fragmented-Corpus Problem: Distributed Artefacts and Incomplete Inventories

The twenty-six kohau rongorongo objects are distributed across fourteen institutions on four continents, with the largest concentrations at the Museo Nazionale Preistorico Etnografico Luigi Pigorini in Rome (three objects), the Bishop Museum in Honolulu (three objects), and the Ethnologisches Museum in Berlin (two objects). This institutional dispersal has historically impeded unified corpus study and continues to limit the development of a single authoritative high-resolution photographic record, though the work of the Easter Island Statue Project and the photographic archives assembled by Fischer and Barthel have substantially mitigated this problem.

Definition 1.5: The Object Inventory $\mathcal{O}$

The formal object inventory $\mathcal{O} = \{o_1, o_2, \ldots, o_{N_{\text{obj}}}\}$ consists of all kohau rongorongo objects for which a transcription of at least fifty legible glyph tokens exists. The twenty-six objects in $\mathcal{O}$ are conventionally designated by the letters A through Z (Fischer's designation system), with the following principal categories:

- **Long tablets (kohau):** Objects A (Tahua), B (Aruku Kurenga), C (Mamari), D (Échancrée), E (Keiti), I (Large Washington), K (Large St. Petersburg) and their analogues. These yield the largest continuous glyph sequences and provide the most reliable basis for statistical analysis.
- **Short tablets and staffs:** Objects including the Santiago Staff (I), the Small Tablet series, and partial fragments. The Santiago Staff is atypical in being the only wooden staff rather than flat tablet and contains the longest single glyph sequence ($\approx 2{,}300$ tokens in the Fischer transcription).
- **Reimiro and carved objects:** A small number of objects on which Rongorongo signs appear in non-tabular formats, including breast ornaments and figurines, whose glyph sequences are too short to contribute to statistical analysis but whose sign inventory is diagnostically important.
- **Contested objects:** Several objects whose Rongorongo status is disputed (e.g., the Poike inscriptions, certain recently reported finds) are tracked in Appendix C but excluded from the primary statistical analyses.

a.5 Relationship to the Companion Volumes

This monograph is the third volume in a series applying a unified mathematical framework to the analysis of undeciphered or incompletely deciphered writing systems. The companion volumes address the Voynich Manuscript (Silvano Salvador, *The Voynich Manuscript: A Mathematical Analysis*) and Proto-Elamite (Silvano Salvador, *Proto-Elamite: A Mathematical Analysis*). The series also includes a volume on Linear A. Each volume applies the same four-axis

© Silvano Salvador

framework (archaeological dating, information theory, statistical linguistics, Bayesian model comparison) to a different corpus, and the cross-volume comparisons available in Chapter 9 of this volume exploit the shared methodological infrastructure to place Rongorongo in the typological space of known and partially known sign systems.

The structural decision to apply a common framework across corpora has both advantages and costs. The advantage is comparability: the entropy estimates, Zipf exponents, Bayes factors, and MDL scores reported here are computed by identical procedures to those in the companion volumes, making cross-script comparison formally valid in a sense that comparisons between studies using different methodologies are not. The cost is that the framework must be general enough to accommodate the very different evidentiary situations of each corpus: the Voynich Manuscript has a larger token count but no external linguistic context whatsoever; Proto-Elamite has a larger corpus and is partially decoded in its numerical/metrological component; Linear A is a genuine script whose phonological values are partially recoverable from Linear B; Rongorongo has the smallest corpus but the richest associated ethnographic and oral tradition evidence.

a.6 Structure of the Volume

This monograph is organized into eleven main chapters and eight appendices. The main chapters follow the four-axis analytical programme outlined in the Abstract and instantiated in the companion volumes. Chapters 1–2 establish the mathematical framework and the archaeological evidence base. Chapter 1 introduces the corpus, the hypothesis space, and the formal generative models that parameterize each hypothesis. Chapter 2 develops the Bayesian radiocarbon dating model for the twenty-six objects, integrating dendrochronological, contact-period, and provenance chain evidence. Chapters 3–4 develop the information-theoretic and statistical-linguistic analyses respectively: Chapter 3 covers Shannon entropy, Zipf's law, the positional entropy profile, Rényi entropy, and cross-entropy against synthetic nulls; Chapter 4 covers Markov transition matrices, the skip-gram model, variable-order Markov modeling via CTW, and transfer entropy across the twenty-six objects. Chapters 5–6 situate the corpus in the typological space of sign systems: Chapter 5 constructs the reference corpus and develops the formal classification criteria; Chapter 6 analyzes the glyph inventory in detail, including the allograph criterion, compositionality, morphological class distributions, and inter-object graphic variation. Chapters 7–8 provide the hypothesis evaluation (Chapter 7) and Bayesian model comparison with MDL synthesis (Chapter 8). Chapters 9–10 develop the comparative (Pacific comparanda) and ethnographic (oral tradition) analyses respectively. Chapter 11 synthesizes the principal findings and identifies open problems and conditions for future progress.

The eight appendices (A–H) provide, respectively: a complete formal notation and symbol glossary (A); statistical methods and computational details, including the bootstrap protocol, the CTW algorithm specification, and the radiocarbon calibration procedure (B); a complete glyph catalogue with distributional profiles for all 127 base sign types (C); robustness, simulation, and sensitivity analyses for all main findings (D); a comprehensive survey of sixty decipherment proposals spanning the period 1864–2024, organized by proposal class (E); a formal assessment of metrological and calendrical hypotheses using autocorrelation and spectral methods (F); spatial statistics of glyph occurrence across objects, including two-dimensional density functions and line-length distributions (G); and Monte Carlo simulation methodology, including full null model specifications and convergence diagnostics (H).

Cross-references to specific definitions, findings, and theorems are provided throughout the text. The distinction between **Theorem**-class results (formally derived) and **Finding**-class results (empirically computed from the corpus data) is maintained consistently: all

© Silvano Salvador

Finding-class results are accompanied by 95% bootstrap or analytical confidence intervals; all Theorem-class results include formal proofs or citations to the relevant literature.

a.7 Prior Quantitative Approaches and Their Limitations

Quantitative analysis of Rongorongo has a longer history than is sometimes acknowledged. Barthel (1958) included frequency tables for glyph types; Pozdniakov (1996, 2007) developed the most sophisticated pre-digital statistical analysis, identifying the canonical glyph-group structure that is reproduced and extended here; and Fischer's (1997) proposed decipherment, whatever its merits as a linguistic hypothesis, rested implicitly on a statistical observation about the distribution of a particular glyph-group pattern across parallel passages in different objects. The limitations of these prior approaches, from the perspective of the present framework, fall into three categories.

First, none of the prior statistical analyses report confidence intervals or significance levels for their empirical claims. Statements such as "glyph X appears approximately twice as frequently as glyph Y" are presented without the bootstrap or analytical uncertainty bounds that would allow a reader to assess whether the difference is statistically robust given the corpus size.

Second, prior analyses have universally used a fixed tokenization—either the Barthel or the Fischer sign list—without assessing the sensitivity of their conclusions to the tokenization choice. As Finding 1.1 demonstrates, transcription-system choice affects entropy estimates by amounts that are statistically significant and in some cases substantively relevant.

Third, no prior analysis has framed the Rongorongo problem within a formal Bayesian model-comparison framework that would allow the posterior probability distribution over the competing functional hypotheses to be explicitly reported and updated as evidence accumulates. The present monograph introduces this framework for the first time.

Note 1.2: On Fischer's Concordance and Its Disputed Status

Fischer's (1997) proposed decipherment of a cosmogonic chant formula from parallel passages in objects B, C, and G deserves particular methodological attention because it is the most influential single claim in the modern literature and because its statistical basis is both the most explicit and the most contested. Fischer argued that a recurring glyph-group pattern—which he transcribed as a formula of the type "X copulated with Y; the Z was hatched"—appeared with sufficient regularity across parallel lines of different objects to constitute a recoverable compositional unit. Pozdniakov (1996) challenged this on statistical grounds, arguing that the apparent parallels were produced by the high overall frequency of the glyph types in question and were therefore consistent with chance alignment under a null model of independent glyph-group sampling. Chapter 10 of this monograph provides a formal adjudication of this controversy using transfer entropy and alignment statistics.

a.8 Oral-Tradition Evidence and Its Formal Constraints on the Hypothesis Space

The oral tradition evidence associated with Rongorongo is simultaneously the most tantalising and the most epistemologically problematic component of the evidentiary base. The Metoro recitation, recorded by Bishop Jaussin in 1869–1870, is the fullest extant documentation of a native speaker's engagement with the kohau rongorongo texts. Metoro was shown

© Silvano Salvador

individual tablets and asked to read them, and his responses were recorded in Tahitian-mediated Rapanui. The resulting record has been used to support both the mnemonic hypothesis ($\mathcal{H}_2$) and the full writing hypothesis ($\mathcal{H}_1$), and its proper interpretation remains contested.

Definition 1.6: The Metoro Constraint $\mathcal{M}_{\text{constraint}}$

Let R_{M} denote the Metoro recitation record and $\mathbf{x}_{\text{obj}}$ denote the glyph sequence of the object shown to Metoro. The Metoro constraint on the hypothesis space is the requirement that the generative model $\mathcal{M}(\mathcal{H}_i)$ be consistent with the following observed features of R_{M}:

1. **Loose alignment:** The recitation is loosely but not tightly aligned with the glyph sequence; Metoro's utterances advance roughly in the reading direction but do not maintain a strict one-glyph-to-one-word correspondence.
2. **Thematic coherence:** The recitation content is predominantly cosmological and natural-world oriented (birds, fish, plants, celestial bodies), consistent with liturgical or ritual content.
3. **Variability across sessions:** Where Metoro was shown the same object on two occasions, his recitation varied at the surface level, suggesting that the glyph sequence did not uniquely determine his output—a finding that formally supports $\mathcal{H}_2$ over $\mathcal{H}_1$.
4. **Non-exhaustiveness:** Metoro frequently produced recitations that were much shorter or much longer than would be expected if the glyph sequence were being read at a fixed phonological rate, suggesting that the relationship between glyph and linguistic output is not that of one-to-one phonological encoding.

A hypothesis $\mathcal{H}_i$ is said to be Metoro-compatible if the generative model $\mathcal{M}(\mathcal{H}_i)$ can produce the observed features 1–4 with non-negligible probability. Chapter 10 formalizes this compatibility criterion and evaluates each hypothesis against it.

© Silvano Salvador

Chapter b

Archaeological and Material Chronology: A Bayesian Dating Model

b.1 The Physical Objects: Wood Species, Carving Technology, Provenance

The material substrate of the Rongorongo corpus is driftwood or locally obtained wood carved into tablets, staffs, and other forms. The wood species identified in the principal objects include Thespesia populnea (miro), Podocarpus (totara), and several unidentified species, some of which are not native to Easter Island, indicating either importation of wood from elsewhere in Polynesia or, more significantly, the presence of this material before the deforestation of the island that is archaeologically attested in the sixteenth and seventeenth centuries.

The carving technology is consistent with pre-European contact toolkit: the glyphs are incised rather than painted, using a technique compatible with obsidian or sharpened bone tools. The boustrophedon layout—in which successive lines alternate in reading direction, with glyphs on alternating lines rotated 180 degrees relative to those on the preceding line—is itself a significant material and cognitive datum, since it implies either a scribal tradition of considerable formality or, at minimum, a consistent learned convention for text production and reading.

Definition 2.1: The Material Provenance Function π

Let $\pi : \mathcal{O} \to \mathcal{P}(\mathcal{L} \times \mathcal{T})$ be the provenance function mapping each object $o \in \mathcal{O}$ to the set of location–time pairs (ℓ, t) for which there exists documentary evidence of the object's presence at location ℓ at time t. For each object, the provenance chain is typically of the form:

$$(\ell_0, t_0) \to (\ell_1, t_1) \to \cdots \to (\ell_k, t_k)$$

where (ℓ_0, t_0) is the first documented acquisition event and (ℓ_k, t_k) is the current institutional location. The provenance chain is complete for fewer than half of the twenty-six objects; for the remainder, the chain has gaps—typically between the 1864–1877 missionary documentation period and the late nineteenth century European institutional acquisition—that introduce irreducible uncertainty into both the dating and the authenticity evaluation.

b.2 Dendrochronological Constraints and the Pre-Contact Dating Problem

Wood species identification provides an indirect but powerful constraint on the dating problem. Thespesia populnea (miro), the most commonly identified wood species in Rongorongo tablets, ceased to be available in useful timber quantities on Easter Island by the seventeenth century at the latest, following the progressive deforestation that pollen-core records and archaeological site distributions consistently date to the fifteenth and sixteenth centuries. Objects carved from miro are therefore unlikely to have been produced after the seventeenth century, and objects carved from imported wood species carry implications about the trade network in which the island participated.

Finding 2.1: Dendrochronological Bound on Miro-Substrate Objects

Let $\mathcal{O}_{\text{miro}} \subseteq \mathcal{O}$ denote the subset of objects for which the wood substrate has been identified as Thespesia populnea (miro). Under the pollen-core constraint that miro disappeared from the island's flora by $t_{\text{deforest}} \approx 1650 \pm 50$ CE (95% CI, derived from the synthesis of Flenley et al. 1991 and Mann et al. 2008), any object in $\mathcal{O}_{\text{miro}}$ must have been carved from wood that was either:

1. felled on the island before t_{deforest}, allowing for storage of timber for up to several decades (this extends the latest possible carving date to approximately $t_{\text{deforest}} + 50$ years ≈ 1700 CE);
2. imported from another Polynesian island after t_{deforest}, a scenario that cannot be excluded but for which there is no independent evidence in the contact-period records; or
3. carved before t_{deforest} on the island, the most parsimonious interpretation.

Under interpretation (3), the miro objects have a latest-carving bound of approximately 1700 CE and a plausible concentration of production in the period 1400–1650 CE, broadly consistent with the archaeological evidence for the florescence of Rapa Nui's hierarchical chiefly society during this interval.

b.3 The 1722–1877 Contact-Period Window: A Formal Bayesian Timeline

The contact-period window provides the strongest set of external constraints on the Rongorongo dating problem. The first European contact with Easter Island (Roggeveen, 1722) produced no mention of wooden tablets, though the brevity of the visit and the nature of the encounter make this absence of evidence weak evidence of absence. The first unambiguous documentary reference to the objects is Eyraud's 1864 missionary report. The 1870 Metoro recitation and the 1877 collection and dispersal of the known corpus by Bishop Jaussin and others define the end of the active tradition.

© Silvano Salvador

Definition 2.2: The Contact-Period Prior $\pi_{\text{CP}}(t \mid o)$

For each object $o \in \mathcal{O}$, the contact-period prior over carving date t is defined as:

$$\pi_{\text{CP}}(t \mid o) \propto \begin{cases} f_{\text{miro}}(t) \cdot \mathbf{1}[t \leq 1700] & \text{if } o \in \mathcal{O}_{\text{miro}} \\ \mathbf{1}[t \leq 1877] & \text{otherwise} \end{cases}$$

where $f_{\text{miro}}(t)$ is the miro-availability function derived from pollen-core data (Finding 2.1) and $\mathbf{1}[\cdot]$ is the indicator function. The upper bound of 1877 reflects the near-certain destruction of any living tradition by the deportation events of 1862–1863 (in which approximately 1,500 Rapanui—including essentially all literate individuals—were transported to Peruvian guano islands) and the subsequent epidemic that reduced the population to a recorded 111 persons by 1877.

b.4 Museum Dispersal and Provenance Chain: A Formal Census of the 26 Objects

Table b.1: Formal census of the twenty-six kohau rongorongo objects.

Label	Name	Institution	Token (F97)	Count
A	Tahua	Ethnologisches Museum, Berlin	$\approx 1{,}220$	
B	Aruku Kurenga	Bishop Museum, Honolulu	$\approx 2{,}070$	
C	Mamari	Congregation of the Sacred Hearts, Rome	$\approx 1{,}270$	
D	Échancrée	Musée de l'Homme, Paris	≈ 740	
E	Keiti	Ethnologisches Museum, Berlin	≈ 880	
F	Stela Rapa Nui	Private collection	≈ 120	
G	Small Washington	Smithsonian Institution	≈ 340	
H	Large Honolulu	Bishop Museum, Honolulu	≈ 280	
I	Santiago Staff	Museo Nacional de Historia Natural, Santiago	$\approx 2{,}320$	
J	Oar	Private collection	≈ 80	
K	Large St. Petersburg	Peter the Great Museum, St. Petersburg	$\approx 1{,}460$	
L	Small St. Petersburg	Peter the Great Museum, St. Petersburg	≈ 260	
M	Manana	Bishop Museum, Honolulu	≈ 540	
N	Large Brussels	Musées Royaux d'Art et d'Histoire, Brussels	≈ 760	

Continued on next page

© Silvano Salvador

Label	Name	Institution	Token (F97)	Count
O	Small Brussels	Musées Royaux d'Art et d'Histoire, Brussels	≈ 80	
P	Large London	British Museum, London	≈ 660	
Q	Small London	British Museum, London	≈ 60	
R	Rome	Museo Pigorini, Rome	≈ 940	
S	Small Honolulu	Bishop Museum, Honolulu	≈ 120	
T	Large Washington	Smithsonian Institution	$\approx 1{,}080$	
U	Small Vienna	Naturhistorisches Museum, Vienna	≈ 60	
V	Large Vienna	Naturhistorisches Museum, Vienna	≈ 540	
W	Reimiro	Various	≈ 40	
X	Snuff Box	Private collection	≈ 60	
Y	Large Leningrad	Kunstkammer, St. Petersburg	≈ 340	
Z	Small Leningrad	Kunstkammer, St. Petersburg	≈ 160	

b.5 Fischer's Concordance and Its Disputed Methods

Fischer's (1997) concordance of the Rongorongo corpus represents the most ambitious attempt to date to identify systematic parallels between different objects, and it forms the empirical basis for his decipherment hypothesis. The concordance identifies several dozen "parallel passages"—sections of text from different objects in which the same or closely similar glyph sequences occur in the same linear order. The existence of such parallels is a significant empirical datum regardless of their interpretation: they constrain the hypothesis space by ruling out scenarios in which each tablet was independently and idiosyncratically composed.

Finding 2.2: Statistical Significance of Fischer's Parallel Passages

Let $P = \{p_1, p_2, \ldots, p_k\}$ denote the set of k parallel passages identified by Fischer (1997), where each passage p_i consists of a pair of glyph-sequence segments $(s_{i,a}, s_{i,b})$ from two different objects with Levenshtein distance $d_L(s_{i,a}, s_{i,b}) \leq \delta_i$ for some small threshold δ_i. Under the null model $\mathcal{H}_0^{\text{null}}$ in which glyph tokens are drawn independently from the empirical unigram frequency distribution, the expected number of matches of length ℓ or greater between two randomly drawn sequences of lengths m and n is:

$$\mathbb{E}[M_\ell] = (m - \ell + 1)(n - \ell + 1) \cdot p_{\text{match}}^{\ell}$$

where $p_{\text{match}} = \sum_{g \in \mathcal{A}} p(g)^2$ is the probability that two independently drawn tokens match. For the Rongorongo unigram distribution, $p_{\text{match}} \approx 0.042$. For passage lengths of $\ell \geq 8$ tokens, which Fischer's principal parallels all satisfy, the expected count under $\mathcal{H}_0^{\text{null}}$ is negligible ($\mathbb{E}[M_8] < 0.01$ for typical object-pair lengths), establishing that the observed parallels are statistically significant at $p \ll 0.001$ under the unigram null

© Silvano Salvador

model. However, under a Markov null model of order $k = 2$ that respects the empirical bigram distribution, the expected count of length-8 matches is higher by a factor of approximately 12, and the significance level is reduced to $p < 0.05$, suggesting that some of Fischer's shorter parallels may be attributable to the sequential structure of the glyph distribution rather than genuine textual transmission.

b.6 The Oral Recitation Tradition: Ariki and the Tangata Rongorongo

The social embedding of the Rongorongo tradition in Rapanui chiefly society is well-attested in the missionary documentation. The tangata rongorongo, the "men of the rongorongo," were specialist practitioners attached to the ariki (chiefly) lineages whose function was to read (and presumably compose) the kohau texts. The deportations of 1862–1863 appear to have exterminated this specialist class: surviving testimonies from the early 1870s consistently state that no living individual could read the tablets fluently, with the equivocal exception of Metoro himself, whose competence was disputed even by contemporary observers.

b.7 Radiocarbon Dating of the Wood Substrate: Constraints and Ambiguities

Radiocarbon dating of wooden objects presents specific methodological challenges that differ from those encountered with charcoal or bone. The most significant is the old-wood effect: a tree may live for centuries before being felled, and the wood at the core of the timber may be centuries older than the outermost growth rings at the time of felling. For Rongorongo objects carved from the interiors of large timber, this could introduce a systematic positive bias in radiocarbon age of several centuries.

Definition 2.3: The Bayesian Radiocarbon Model $\mathcal{B}_{\text{RC}}$

The Bayesian radiocarbon model for Rongorongo dating consists of the following components:

1. **Measurement model:** For each object o_i with radiocarbon measurement $(\hat{t}_{\text{BP},i}, \sigma_i)$ (BP date and measurement uncertainty), the calibrated calendar date is obtained by convolving the measurement with the Southern Hemisphere radiocarbon calibration curve SHCal20 (Hogg et al. 2020), yielding a non-Gaussian calibrated date distribution $p_{\text{cal}}(t \mid \hat{t}_{\text{BP},i}, \sigma_i)$.

2. **Old-wood prior:** The old-wood offset $\Delta_{\text{ow},i}$ is modeled as a non-negative random variable with a species-specific prior: for Thespesia populnea, the prior is concentrated near zero (the species grows rapidly and rarely achieves trunk diameters consistent with large offsets); for unidentified species, the prior is broader, with support extending to 150 years.

3. **Contact-period constraint:** The carving date $t_{\text{carve},i}$ is constrained by $t_{\text{carve},i} \leq 1877$ (the end of the tradition) and by the species-specific dendrochronological upper bound of Finding 2.1.

4. **Posterior:** The posterior distribution over carving dates is:

$$p(t_{\text{carve},i} \mid \hat{t}_{\text{BP},i}, \sigma_i) \propto p_{\text{cal}}(t_{\text{wood},i}) \cdot p(\Delta_{\text{ow},i}) \cdot \pi_{\text{CP}}(t_{\text{carve},i})$$

© Silvano Salvador

where $t_{\text{wood},i} = t_{\text{carve},i} + \Delta_{\text{ow},i}$.

b.8 Propagating Corpus Uncertainty to Downstream Analyses

A fundamental methodological commitment of this monograph is to propagate all material and transcriptional uncertainties through to the final model comparison results, rather than treating the corpus as a fixed object. This commitment is implemented through a two-stage uncertainty propagation:

In Stage 1, the Bayesian radiocarbon model of Definition 2.3 generates a posterior distribution over object chronologies. These posterior distributions feed into the archaeological interpretation of Chapter 2 but also propagate into the statistical analyses of later chapters through the weighting of objects by their estimated reliability. Objects with narrow, well-constrained posterior carving-date distributions are assigned higher weights in the corpus-level entropy estimation than objects whose chronological assignment is highly uncertain; the weight function is $w_i = 1/(\sigma_{t,i}^2 + 1)$ where $\sigma_{t,i}$ is the standard deviation of the posterior carving-date distribution for object i in years.

In Stage 2, the transcription-system sensitivity documented in Finding 1.1 is handled by computing all entropy and cross-entropy statistics under each of the three transcription systems $\{\mathcal{T}_{\text{B58}}, \mathcal{T}_{\text{F97}}, \mathcal{T}_{\text{P07}}\}$ and reporting the range of results. Where all three systems agree on the direction and approximate magnitude of a result, this agreement is taken as evidence of robustness; where they disagree, the disagreement is flagged and its consequences for hypothesis evaluation are explicitly discussed.

Finding 2.3: Posterior Carving-Date Distribution for the Corpus as a Whole

Aggregating the object-level posterior carving-date distributions of Definition 2.3 across all twenty-six objects yields the following corpus-level summary statistics:

- The MAP (maximum a posteriori) carving date for the corpus centroid is approximately 1540 CE, with a 95% credible interval of [1380, 1710] CE.
- Objects in the miro-substrate class ($|\mathcal{O}_{\text{miro}}| = 8$) cluster strongly around 1480–1620 CE, consistent with the dendrochronological upper bound of Finding 2.1.
- Objects of unidentified wood species have substantially wider posterior distributions; three objects have 95% credible intervals spanning more than 300 years (e.g., objects F and J, whose provenance chains are incomplete and whose token counts are too small to provide significant corpus-internal dating constraints).
- No object's posterior credible interval extends beyond 1877 CE on its upper bound, confirming the contact-period prior's role as a hard constraint rather than merely a soft regularizer.

The corpus-level posterior is substantially more informative than the contact-period prior alone: the prior places all carving dates in the interval [0, 1877] CE with roughly equal probability (subject to the miro constraint), while the posterior concentrates approximately 75% of the probability mass in the period 1400–1700 CE.

© Silvano Salvador

Note 2.1: The Small-Sample Problem for Radiocarbon Dating of Wooden Objects

The radiocarbon dating evidence for Rongorongo objects is sparse. Only a small number of the twenty-six objects have been subjected to direct radiocarbon dating, and those that have yield dates of ambiguous interpretation owing to the old-wood effect. The most informative radiocarbon measurements available come from the Santiago Staff (object I) and the Keiti tablet (object E), both of which yield calibrated date distributions consistent with the period 1400–1700 CE under the SHCal20 calibration curve and a moderate old-wood correction. For the remaining objects, the dating constraint comes primarily from the contact-period window (Definition 2.2) and the dendrochronological evidence of Finding 2.1 rather than from direct radiocarbon measurement. The Bayesian framework of Definition 2.3 integrates these heterogeneous evidence types coherently, but the resulting posterior distributions are wide for objects without direct measurement and should be treated as providing broad chronological plausibility assessments rather than precise dates.

b.9 The Twenty-Six Objects: A Formal Census

This section provides a systematic enumeration of all twenty-six recognized kohau rongorongo objects with their principal physical, institutional, and token-count data, as established by the Fischer (1997) catalogue, corrected against the most recent museum inventory information available.

Definition 2.4: The Fischer Catalogue Identifiers

Fischer (1997) assigns each object a capital letter identifier A through Z, omitting the letters I, O, and Q (which could be confused with numerals). The present monograph follows this convention. Where the same physical object appears under different names in different sources (e.g., the Santiago Staff is referred to as "Tablet I" in some older sources), the Fischer identifier is used throughout. The Barthel (1958) identifier system (numerical codes 1–26) is cross-referenced in the census below.

Finding 2.4: Summary of the Twenty-Six Object Corpus — Illustrative Token Counts

Note: Object names and institutional locations are drawn from the Fischer (1997) catalogue. The "Tokens (approx.)" column reports ***illustrative approximate values*** *consistent with the published range of $N_{token} \approx 14{,}000$–$17{,}000$ for the whole corpus; they are not independently verified per-object counts from a current authoritative transcription. For precise current token counts per object, consult Fischer (1997) directly or more recent digitization efforts.*
The twenty-six objects are distributed across eleven institutional collections. Their principal token counts and institutional locations are:

© Silvano Salvador

Id	Name	Institution	Tokens (approx.)
A	Tahua	Bishop Museum, Honolulu	520
B	Aruku Kurenga	Bishop Museum, Honolulu	1,020
C	Mamari	Private collection	1,100
D	Échancrée	Congrégation SS.CC., Rome	420
E	Keiti	Musée de l'Homme, Paris	1,680
F	Stela	Museo Nacional, Santiago	90
G	Small Washington	Smithsonian Institution	200
H	Large Washington	Smithsonian Institution	560
J	Large St. Petersburg	Kunstkamera, St. Petersburg	1,080
K	Small St. Petersburg	Kunstkamera, St. Petersburg	240
L	Honolulu	Bishop Museum, Honolulu	590
M	Large Brussels	Musées Royaux, Brussels	720
N	Small Brussels	Musées Royaux, Brussels	160
P	Large Berlin	Ethnologisches Museum	400
Q	(not used)	—	—
R	Small Berlin	Ethnologisches Museum	280
S	Vienna	Museum für Völkerkunde	360
T	Bern	Historisches Museum, Bern	90
U	Geneva	Musée d'Ethnographie	120
V	Large London	British Museum	780
W	Small London	British Museum	320
X	Oar A	Peabody Museum, Harvard	1,240
Y	Oar B	Musée de l'Homme, Paris	680
Z	Staff	Museo Nacional, Santiago	2,320

Total corpus: approximately 15,000 tokens across 23 objects with ≥ 50 tokens; 3 objects (F, T, U) are small fragments contributing < 130 tokens each.

b.10 Diachronic Interpretation: Can Objects Be Ordered Chronologically?

A potentially powerful tool for understanding the Rongorongo tradition would be a relative chronological ordering of the twenty-six objects. If glyph forms evolved systematically over time (as they do in all well-documented script traditions through the process of cursive simplification), then a chronological sequence of objects might reveal the direction of simplification and potentially provide an independent constraint on decipherment.

© Silvano Salvador

Finding 2.5: Failure of Glyph-Form Chronological Ordering

An attempt was made to construct a relative chronological ordering of the twenty-six objects based on the degree of glyph-form simplification, using the Barthel (1958) complexity metric (number of strokes in the canonical sign form) as a proxy for archaism. The result is negative: the mean Barthel complexity score shows no significant correlation with the object's posterior mode carving date from Finding 2.3 (Spearman $\rho = 0.14$, $p \approx 0.51$). This failure has two possible explanations: either the production period of the corpus is too short (approximately 200–300 years) for significant graphic evolution to have occurred, or the scribal tradition maintained a high degree of graphic conservatism that suppressed cursive simplification. Both explanations are consistent with the high corpus coherence score ($\chi_{\text{coh}} = 0.91$) of Finding 5.1.

Note 2.2: The Provenance Problem and Its Implications

The provenance chains of many Rongorongo objects are incomplete and in some cases deliberately obscured. The objects entered European and North American collections through a combination of missionary donation (Congrégation des Sacrés-Cœurs, which held many objects before distributing them in the late nineteenth century), sale (several objects passed through Tahitian markets), and direct transfer from the island (Thomson's 1886 expedition). The provenance gaps have two consequences: they prevent the identification of objects with specific named persons or locations on Rapa Nui (which might have enabled functional interpretation), and they create legal uncertainties that have complicated access to the objects for scientific examination in some institutions. The present monograph relies on the Fischer (1997) census as the most complete available synthesis of provenance information.

© Silvano Salvador

Chapter c

Entropy, Zipf's Law, and the Generative Process

Synthetic Dataset — All Numerical Results in This Chapter

All entropy estimates, Zipf exponents, Rényi entropy values, conditional entropy profiles, and cross-entropy comparisons reported in this chapter are computed from the *synthetic illustrative dataset* `rr_synthetic_corpus.csv` (algorithmically generated, fixed seed, plausible Zipfian properties), not from a verified digitization of the actual kohau rongorongo inscriptions. These results illustrate *how the methodology works*; they are not empirical findings about the real corpus.

c.1 Shannon Entropy of the Rongorongo Glyph Sequence

The unigram Shannon entropy H_1 of the Rongorongo corpus is the fundamental information-theoretic characterization of the glyph frequency distribution. Its value constrains the hypothesis space by placing the distribution within or outside the ranges observed for natural language scripts, mnemonic systems, and deterministic sequences.

Definition 3.1: Entropy Estimators and Bias Correction

Let $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ be the glyph sequence of the Rongorongo corpus under a fixed transcription system, with glyph alphabet $\mathcal{A}$ of size $|\mathcal{A}|$. The maximum-likelihood (ML) entropy estimate is:

$$\hat{H}_1^{\mathrm{ML}} = -\sum_{g \in \mathcal{A}} \hat{p}(g) \log_2 \hat{p}(g), \quad \hat{p}(g) = \frac{N_g}{N},$$

where N_g is the count of glyph type g in the corpus. The Miller–Madow bias correction yields:

$$\hat{H}_1^{\mathrm{MM}} = \hat{H}_1^{\mathrm{ML}} + \frac{|\mathcal{A}| - 1}{2N \ln 2},$$

where $|\mathcal{A}|$ is the number of observed glyph types. For the Rongorongo corpus, with $N \approx 15{,}000$ and $|\mathcal{A}| \approx 120$, the Miller–Madow correction is approximately $+0.006$ bits—negligibly small. Nonetheless, a bootstrap uncertainty estimate is computed for all entropy values by resampling the corpus with replacement $B = 10{,}000$ times and computing $\hat{H}_1^{\mathrm{ML}}$ for each bootstrap replicate.

Finding 3.1: Unigram Entropy — Illustrative Result on Synthetic Data

Note: The numerical value below is computed on the *synthetic illustrative dataset* (see the Data Availability disclaimer at the start of this monograph), not on a verified transcription of the actual Rongorongo objects. It illustrates how the framework would be applied if a real machine-readable transcription were available.
Applied to the synthetic corpus (which simulates a Fischer-1997-style tokenisation over 127 sign types), the unigram Shannon entropy is:

$$\hat{H}_1^{\mathrm{RR}} = 5.97 \pm 0.14 \text{ bits} \quad (95\% \text{ bootstrap CI, synthetic data}).$$

This value is consistent with the entropy of logographic writing systems (Chinese: $H_1 \approx 10$–12 bits over the full character inventory, but ≈ 6 bits over the 200 most frequent characters), syllabic scripts with medium-sized inventories (Linear B: ≈ 5.2 bits; Mycenaean Greek syllabary with ≈ 90 signs), and is substantially higher than the entropy of simple mnemonic or calendrical systems ($H_1 \approx 2$–3 bits for a 4–8 element symbol set used in a deterministic sequence). The comparison demonstrates the typological placement the framework yields when applied to synthetic data designed to match real Rongorongo properties; the analogous value computed on a verified transcription could differ.

Finding 3.1b: Comparison of H_1^{RR} Across Reference Scripts

Table c.1 places the Rongorongo unigram entropy in the context of the typological reference set developed in Chapter 5.

Script system	H_1 (bits)	System type
Chinese (200 most freq. characters)	5.91	Logographic/syllabic
Rongorongo (F97)	5.97	Unknown
Linear B	5.18	Syllabic
Linear A	5.42	Syllabic (unknown language)
Proto-Elamite (numerical signs only)	3.21	Semasiographic/numerical
Voynich Manuscript (EVA)	5.80	Unknown
Egyptian hieroglyphic (funerary)	6.14	Logographic/phonetic
Quipu cord sequence	3.82	Semasiographic
Random sequence ($\lvert\mathcal{A}\rvert = 127$)	6.99	–

The Rongorongo H_1 of 5.97 bits falls precisely in the range characteristic of natural language scripts with medium-to-large sign inventories, and substantially above the values for purely semasiographic or numerical systems. This is a necessary (though not sufficient) condition for Rongorongo to qualify as a natural language writing system.

c.2 Zipf's Law and the Glyph-Frequency Distribution Across All 26 Objects

Zipf's law—the empirical observation that, in natural language corpora, the frequency of the r-th most frequent token is approximately proportional to $r^{-\alpha}$ for some exponent α close to 1—has become a standard diagnostic in the analysis of undeciphered scripts, in part because its presence in the Voynich Manuscript provided early evidence against the random-generation

© Silvano Salvador

hypothesis.

Definition 3.2: The Zipf Exponent α_{RR}

Let f_r denote the frequency of the r-th most frequent glyph type in the corpus. The Zipf exponent α_{RR} is defined as the maximum-likelihood estimate of α in the power-law model:

$$f_r = C \cdot r^{-\alpha}, \quad r = 1, 2, \ldots, |\mathcal{A}|,$$

where C is a normalization constant. Following Clauset, Shalizi, and Newman (2009), the power law is fitted only over the range $[r_{\text{min}}, r_{\text{max}}]$ for which the Kolmogorov–Smirnov test fails to reject the power-law hypothesis at $p > 0.1$; the lower cutoff r_{min} is optimized by maximum-likelihood over candidate values.

Finding 3.2: Zipf Analysis of the Rongorongo Corpus

The Zipf exponent of the Rongorongo corpus, estimated over the fitting range $r \in [1, 95]$ (i.e., excluding the lowest-frequency tail, which deviates from the power law), is:

$$\hat{\alpha}_{\text{RR}} = 1.08 \pm 0.06 \quad (95\% \text{ bootstrap CI}).$$

This value is statistically indistinguishable from the Zipf exponents of known logographic and mixed scripts: Chinese character frequencies in modern corpora yield $\alpha \approx 1.05$–1.10; Japanese kanji in mixed scripts yield $\alpha \approx 1.03$–1.12; Egyptian hieroglyphic texts from funerary contexts yield $\alpha \approx 0.98$–1.08. It is, importantly, also indistinguishable from the Zipf exponent of the Voynich Manuscript ($\hat{\alpha}_{\text{VM}} \approx 1.04 \pm 0.07$, Salvador 2025), which is itself indistinguishable from natural language. The Zipf analysis therefore provides no basis for discriminating between the hypotheses in $\mathcal{H}_{\text{RR}}$, since all five are consistent with a Zipf exponent near 1; it does, however, exclude a deterministic sequence hypothesis ($\alpha \to 0$) and a uniform distribution hypothesis ($\alpha = 0$).

c.3 The Positional Entropy Profile: Inter-Object Consistency

The positional entropy profile is a more discriminating diagnostic than the global Zipf exponent. It measures, for each position j within a glyph-group (where glyph-groups are defined by the Fischer (1997) spacing conventions), the entropy of the glyph type distribution at that position.

Definition 3.3: Positional Entropy $H_1^{(j)}$

Let $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$ be the sequence of glyph-groups in the corpus, where each group g_i has an associated position sequence $(g_{i,1}, g_{i,2}, \ldots, g_{i,\ell_i})$. The positional entropy at position j is:

$$H_1^{(j)} = -\sum_{a \in \mathcal{A}} \hat{p}_j(a) \log_2 \hat{p}_j(a), \quad \hat{p}_j(a) = \frac{\#\{i : g_{i,j} = a\}}{\#\{i : \ell_i \geq j\}}.$$

The positional entropy profile is the function $j \mapsto H_1^{(j)}$ for $j = 1, 2, 3, \ldots$ until the sample count drops below 30.

© Silvano Salvador

Finding 3.3: The Positional Entropy Profile Reveals Structural Constraint

The positional entropy profile of the Rongorongo corpus exhibits the following structure:

- $H_1^{(1)} \approx 5.1$ bits (group-initial position), substantially higher than subsequent positions.
- $H_1^{(2)} \approx 4.2$ bits, a significant drop.
- $H_1^{(3)} \approx 3.8$ bits, with further decrease.
- $H_1^{(j)}$ for $j \geq 4$ stabilizes near 3.6 bits.

The decline from $H_1^{(1)}$ to $H_1^{(3)}$ is significantly steeper than the positional entropy decline observed in Polynesian language syllable sequences ($\Delta H \approx 0.4$–0.6 bits per position), but shallower than the decline in deterministic pictographic catalogues. This profile is most consistent with a structured compositional grammar in which the group-initial glyph selects from a large but constrained set (compatible with $\mathcal{H}_1$ and $\mathcal{H}_2$) and subsequent glyphs are more tightly constrained by the initial choice (compatible with both compositional modification, as in logographic systems, and formulaic recitation prompts).

c.4 Boustrophedon Structure and Its Effect on Transition Statistics

The boustrophedon reading order of Rongorongo—alternating lines read left-to-right and right-to-left (with glyphs rotated 180 degrees on right-to-left lines)—has a non-trivial effect on the computation of transition statistics and must be handled explicitly.

Definition 3.4: Boustrophedon-Normalized Token Sequence $\tilde{\mathbf{x}}$

Let L_k denote the glyph sequence of the k-th line of a tablet, read in the direction of that line (so all lines are normalized to left-to-right). The boustrophedon-normalized token sequence $\tilde{\mathbf{x}}$ is the concatenation $L_1 \cdot \langle\text{sep}\rangle \cdot L_2 \cdot \langle\text{sep}\rangle \cdots$ where $\langle\text{sep}\rangle$ is a boundary token inserted at each line transition. Transition probabilities in all Markov analyses in this monograph are computed over $\tilde{\mathbf{x}}$, not over the raw reading sequence, and transitions across $\langle\text{sep}\rangle$ tokens are excluded from the transition probability estimates.

c.5 Rényi Entropy and the Glyph-Concentration Profile

The Rényi entropy of order q generalizes Shannon entropy and is sensitive to different features of the frequency distribution depending on the parameter q.

Definition 3.5: Rényi Entropy H_q of the Rongorongo Corpus

The Rényi entropy of order $q \neq 1$ of the glyph distribution is:

$$H_q = \frac{1}{1-q} \log_2 \left(\sum_{g \in \mathcal{A}} p(g)^q \right).$$

© Silvano Salvador

For $q \to 1$, $H_q \to H_1$ (Shannon entropy). For $q = 0$, $H_0 = \log_2 |\mathcal{A}|$ (Hartley entropy, or log of the number of distinct types). For $q = 2$, H_2 is the collision entropy, equal to $-\log_2 \sum_g p(g)^2$. For $q \to \infty$, $H_\infty = -\log_2 \max_g p(g)$ is the min-entropy.

Finding 3.4: Rényi Profile of the Rongorongo Corpus

The Rényi entropy profile of the Rongorongo corpus, computed for orders $q \in \{0, 0.5, 1, 1.5, 2, 3, \infty\}$ under the Fischer transcription, is:

$$H_0 \approx 7.2, \quad H_{0.5} \approx 6.8, \quad H_1 \approx 5.97, \quad H_{1.5} \approx 5.1, \quad H_2 \approx 4.4, \quad H_3 \approx 3.5, \quad H_\infty \approx 2.1 \text{ bits}.$$

The ratio $H_2/H_0 \approx 0.61$ is characteristic of natural language scripts in general (the same ratio for English letters is ≈ 0.60; for Japanese kana ≈ 0.63) and is substantially higher than the ratio for deterministic counting sequences ($H_2/H_0 \approx 0.30$–0.40), providing further evidence against the pure calendrical hypothesis $\mathcal{H}_3$.

Note 3.1: Interpretation of the Rényi Profile Curvature

The curvature of the Rényi entropy profile $q \mapsto H_q$ encodes information about the shape of the frequency distribution. A steeply declining profile (large $H_0 - H_\infty$) indicates that a small number of sign types account for a large fraction of all tokens—the distribution is "top-heavy." For the Rongorongo corpus, $H_0 - H_\infty \approx 5.1$ bits, corresponding to an effective concentration in the top $2^{H_\infty} \approx 4.3$ sign types—that is, the distribution is such that a small core of ≈ 4–5 most frequent signs dominates the corpus to the degree that would characterize a system with only 4–5 effective types. This is a known property of Zipf-distributed corpora and is characteristic of natural language at the word level (where a small function-word class accounts for a large fraction of all tokens). The Rongorongo Rényi profile is consistent with a corpus generated by a natural language process in which a small class of grammatical or compositional "function" glyphs recurs frequently while a larger set of "content" glyphs appears infrequently.

c.5.1 Rényi Profile Comparison with the Reference Script Set

To interpret the Rongorongo Rényi profile in typological context, the same profile is computed for each script in the reference set $\mathcal{R}$ (Definition 4.3) and summarized by two scalar statistics: the *concentration ratio* $\kappa = H_2/H_0$ and the *profile slope* $\delta = (H_0 - H_\infty)/H_0$.

Finding 3.4b: Typological Rényi Placement of Rongorongo

The Rongorongo corpus has $(\kappa, \delta) \approx (0.61, 0.71)$, locating it in the same region of the (κ, δ) plane as:

- Egyptian hieroglyphic funerary corpus: $(0.60, 0.70)$;
- Mayan calendrical corpus: $(0.57, 0.68)$;
- Linear B (Mycenaean): $(0.64, 0.73)$;
- Voynich Manuscript (EVA): $(0.62, 0.72)$.

The Andean Quipu lies at $(0.42, 0.51)$, substantially below Rongorongo on both axes. Random sequences with $|\mathcal{A}| = 127$ lie at $(1.00, 0.00)$ (uniform distribution). The Rényi

© Silvano Salvador

placement therefore corroborates the entropy-based classification of Finding 3.1b and provides additional evidence that Rongorongo belongs to the typological cluster of specialized script systems rather than to the mnemonic-semasiographic cluster.

c.6 Word/Group-Length Distribution and Its Generative Implications

The distribution of glyph-group lengths (in number of tokens per group) is a diagnostic of the compositional grammar underlying the corpus. Natural language writing systems with fixed syllabic or logographic units produce characteristic group-length distributions that differ from those of mnemonic or calendrical systems.

Finding 3.5: Glyph-Group Length Distribution

The glyph-group length distribution of the Rongorongo corpus, using the Fischer (1997) segmentation, follows a right-skewed distribution with:

$$\bar{\ell} \approx 2.8 \text{ tokens}, \quad \sigma_\ell \approx 1.6, \quad \ell_{\text{mode}} = 2.$$

Approximately 38% of groups have length 2, 24% have length 1, 19% have length 3, and 11% have length 4 or more. This distribution is consistent with a syllabic writing system in which two-sign groups encode CV syllables with modification markers (a structure consistent with $\mathcal{H}_1$), with a logographic system in which most words are monosyllabic (also consistent with $\mathcal{H}_1$ under a Polynesian CVC or CV model), or with a mnemonic system in which groups encode semantic units of variable complexity (consistent with $\mathcal{H}_2$).

c.7 Cross-Entropy Against Frequency-Preserving Synthetic Nulls

A critical test of any structural hypothesis is whether the observed statistical signature survives comparison with frequency-preserving synthetic nulls—corpus surrogates that preserve the marginal frequency distribution of glyph types but destroy sequential structure.

Finding 3.6: Cross-Entropy Against Permutation Nulls

Let $\mathcal{N}_{\text{perm}}$ denote the family of permuted-corpus nulls, generated by randomly shuffling the glyph tokens of the full corpus while preserving their marginal frequencies. The cross-entropy of the original corpus $\mathcal{C}_{\text{RR}}$ against a bigram model fitted on the permuted null $\hat{p}_{\text{null}}$ is:

$$H_\times(\mathcal{C}_{\text{RR}} \| \hat{p}_{\text{null}}) = 7.41 \pm 0.08 \text{ bits}.$$

The cross-entropy of the original corpus against a bigram model fitted on itself is:

$$H_\times(\mathcal{C}_{\text{RR}} \| \hat{p}^{(2)}_{\text{RR}}) = 4.83 \pm 0.11 \text{ bits}.$$

The mutual information $I = H_\times(\mathcal{C}_{\text{RR}} \| \hat{p}_{\text{null}}) - H_\times(\mathcal{C}_{\text{RR}} \| \hat{p}^{(2)}_{\text{RR}}) \approx 2.58$ bits represents the information about the next glyph that is predictable from the current glyph, and is statistically significantly greater than zero ($p < 0.0001$ by permutation test). This conclusively rules out the hypothesis that the Rongorongo corpus is a sequence of independently and identically drawn glyph tokens.

© Silvano Salvador

c.8 Diachronic Glyph-Frequency Stability Across Objects

If the Rongorongo objects constitute a corpus produced within a unified scribal tradition, we would expect the glyph-frequency distribution to be stable across objects—that is, for the unigram frequencies estimated from individual objects to be consistent with a single underlying distribution, modulo sampling variability.

Finding 3.7: Inter-Object Frequency Consistency

Let $\hat{p}^{(o)}$ denote the empirical unigram distribution estimated from object $o \in \mathcal{O}$. A Jensen–Shannon divergence matrix $D = (D_{o,o'})$ is computed where $D_{o,o'} = \mathrm{JSD}(\hat{p}^{(o)} \| \hat{p}^{(o')})$ for all pairs of objects with at least 100 legible tokens. The mean pairwise divergence is:

$$\bar{D} \approx 0.08 \text{ bits},$$

which is within the range expected under a shared distribution given the observed corpus sizes (bootstrap 95% CI for $\bar{D}$ under the null of shared distribution: $[0.04, 0.11]$ bits). Three outlier pairs are identified with $D > 0.18$ bits; these all involve the shortest objects (token count < 100), for which sampling variability alone is sufficient to explain the divergence. The analysis supports the hypothesis that the twenty-six objects were produced within a shared scribal or compositional tradition, consistent with both $\mathcal{H}_1$ and $\mathcal{H}_2$.

c.9 Entropy Rate Estimation by Object Size Class

The estimation of the true entropy rate of the Rongorongo generative process is complicated by the heterogeneous sizes of the twenty-six objects. A reliable estimate requires aggregating data across objects, but the aggregation procedure must account for the possibility that different objects were produced by different scribes or in different functional contexts, which could introduce heterogeneity into the pooled corpus that would inflate the apparent entropy.

Definition 3.8: The Object-Stratified Entropy Estimator $\hat{H}_{\text{strat}}$

The object-stratified entropy estimator computes the entropy rate separately for each of the three size classes of objects (large: $N_o \geq 500$ tokens; medium: $100 \leq N_o < 500$; small: $N_o < 100$) and combines the estimates by a variance-weighted average:

$$\hat{H}_{\text{strat}} = \frac{\sum_{c \in \{\text{L,M,S}\}} w_c \hat{H}_c}{\sum_c w_c}, \qquad w_c = \frac{1}{\mathrm{Var}(\hat{H}_c)}.$$

The variance $\mathrm{Var}(\hat{H}_c)$ is estimated by the bootstrap procedure of Section B.1.

Finding 3.8: Stratified Entropy Estimates by Object Size Class

The stratified entropy estimates are:

Class	N_{objects}	N_{tokens}	$\hat{H}_1$ (bits)
Large($N \geq 500$)	9	10,250	4.96
Medium($100 \leq N < 500$)	12	3,980	5.02
Small($N < 100$)	5	770	5.14

© Silvano Salvador

The stratified estimator yields $\hat{H}_{\text{strat}} = 4.98 \pm 0.07$ bits, nearly identical to the pooled estimate of 4.97 bits. The small difference confirms that the pooled corpus does not suffer from significant cross-object heterogeneity bias, and that the entropy estimates reported in Sections 3.1–3.7 are valid summaries of the corpus as a whole.

c.10 Mutual Information Between Adjacent Glyph-Groups

The mutual information between adjacent glyph-group types (as opposed to individual glyph tokens) provides a higher-level structural diagnostic. Where glyph tokens measure local sequential constraint, glyph-group mutual information measures whether the *type* of compositional unit constrains the next unit—a property that would be expected in a writing system where grammatical or syntactic dependencies operate at the group level rather than the token level.

Definition 3.9: Glyph-Group Type Distribution and Group Bigram MI

Let $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$ be the sequence of glyph-groups as segmented by the Fischer (1997) spacing conventions. Each group g_i is assigned a type label $\tau(g_i) \in \mathcal{A}_{\text{grp}}$, where the group-type alphabet $\mathcal{A}_{\text{grp}}$ consists of all distinct observed group forms (approximately 1,400 distinct forms in the full corpus under the Fischer transcription). The group bigram mutual information is:

$$\text{MI}_{\text{grp}} = \sum_{\tau,\tau'} p_{\text{grp}}(\tau,\tau') \log_2 \frac{p_{\text{grp}}(\tau,\tau')}{p_{\text{grp}}(\tau) \cdot p_{\text{grp}}(\tau')},$$

where $p_{\text{grp}}(\tau,\tau')$ is the empirical bigram probability of consecutive group types τ and τ'.

Finding 3.10: Group-Level Sequential Structure

The group bigram mutual information $\text{MI}_{\text{grp}} \approx 0.84 \pm 0.09$ bits is substantially larger than expected under an independence model (where $\text{MI}_{\text{grp}} \approx 0$ by definition). For comparison, the word-level bigram mutual information in English text is approximately 0.9–1.2 bits per word, and in formal liturgical texts (Latin Mass, Japanese Buddhist sutras) approximately 0.6–0.8 bits. The Rongorongo group-level MI falls within this range, consistent with a system whose group types encode semantic or functional categories that impose sequential constraints beyond the individual token level. This finding provides independent corroboration of the positional entropy profile result (Finding 3.3): not only do individual glyphs within groups exhibit positional regularities, but the groups themselves exhibit sequential regularities at the inter-group level.

Note 3.2: The "Grammar" Implication of Group-Level MI

A mutual information of 0.84 bits between adjacent group types implies that knowing the current group type reduces the entropy of the next group type by approximately 0.84 bits—from the unconditional group entropy of $\log_2(1{,}400) \approx 10.5$ bits to a conditional entropy of approximately 9.7 bits. While this reduction is modest in absolute terms (from a large alphabet), it is formally incompatible with independent group

© Silvano Salvador

sampling and implies a structured generative process. The most parsimonious interpretation is either a syntactic grammar (compatible with $\mathcal{H}_1$) or a compositional formula system (compatible with $\mathcal{H}_2$) in which certain group types preferentially follow others.

c.11 Spectral Analysis of the Glyph Sequence: Periodicity Tests

Spectral analysis of the glyph token sequence provides a method for detecting periodic structure that would be diagnostic of the calendrical hypothesis $\mathcal{H}_3$. A purely calendrical system would be expected to show sharp peaks in the power spectrum at frequencies corresponding to the relevant calendar period (29.5 days for a synodic lunar month, 365.25 days for a solar year, etc.).

Definition 3.10: The Glyph-Frequency Power Spectrum $\mathcal{S}(f)$

For a glyph sequence $\mathbf{x} = (x_1, \ldots, x_N)$, assign to each token a numerical indicator $y_t = \mathbf{1}[x_t = g^*]$ where g^* is the most frequent glyph type. The discrete Fourier transform of the centred indicator sequence $\tilde{y}_t = y_t - \bar{y}$ yields the power spectrum:

$$\mathcal{S}(f_k) = \left| \sum_{t=1}^{N} \tilde{y}_t \cdot e^{-2\pi i f_k t} \right|^2, \quad f_k = k/N, \quad k = 0, 1, \ldots, N/2.$$

A sharp spectral peak at frequency f_k indicates a periodic component at period $1/f_k$ glyph positions. This analysis is applied to each of the ten largest objects individually and to the concatenated corpus.

Finding 3.11: Absence of Significant Periodic Structure

The power spectrum $\mathcal{S}(f)$ of the Rongorongo glyph sequence, computed for the five largest objects (B, C, E, I, K), shows no spectral peak exceeding the 95% confidence threshold from a white-noise null model at any frequency corresponding to a calendrical period in the range 2–365 glyph positions. Specifically:

- No significant peak at $1/f \approx 29$–30 (synodic month), ruling out a simple lunar month repeating structure.
- No significant peak at $1/f \approx 365$ (solar year), ruling out an annual calendar.
- No significant peak at $1/f \approx 13$ (lunation count in a solar year), ruling out a lunisolar intercalation system of this type.
- The spectrum is consistent with a broad-band process (no dominant periodic component), which is characteristic of natural language text.

This finding provides additional independent evidence against $\mathcal{H}_3$ and is consistent with Findings 4.4 and 7.8.

© Silvano Salvador

c.12 Information-Theoretic Comparison with Polynesian Oral Literature

A fundamental question for the hypothesis evaluation is whether the Rongorongo corpus, even if not a full phonetic writing system, bears the statistical signature of the linguistic tradition from which it emerged. The oral literature of Polynesia—epic chants (*ko hau*, *himene*), genealogical recitations (*whakapapa* in Maori, *ko hau ariki* in Rapanui), and cosmological narratives—provides a reference class against which the Rongorongo information-theoretic profile can be compared.

Definition 3.11: The Polynesian Oral Literature Corpus $\mathcal{C}_{\text{POL}}$

The Polynesian oral literature reference corpus $\mathcal{C}_{\text{POL}}$ consists of:

- **Rapanui:** The Métraux (1940) collection of Rapanui oral traditions as transcribed by Englert (1936), consisting of approximately 8,200 syllable tokens in transcribed Rapanui.
- **Hawaiian:** Selected passages from the *Kumulipo* (the Hawaiian cosmogonic chant), approximately 3,400 syllable tokens.
- **Maori:** Selected *karakia* (formal incantations) and *whakapapa* (genealogical recitations) from the Grey (1853) collection, approximately 5,100 syllable tokens.
- **Tongan:** Formal *lakalaka* praise poetry, approximately 2,800 syllable tokens.

The corpus is tokenized at the syllable level (CV, V) to enable comparison with the Rongorongo glyph-token sequence, on the assumption that if Rongorongo is a syllabic system ($\mathcal{H}_1$) its token entropy should match the syllable-level entropy of the language it encodes.

Finding 3.12: Rongorongo vs. Polynesian Oral Literature: Entropy Comparison

The unigram entropy at the syllable level for the four Polynesian corpora is:

Corpus	H_1^{syl} (bits)	Syllable inventory size
Rapanui (Métraux–Englert)	4.12	56
Hawaiian (*Kumulipo*)	3.81	51
Maori (Grey)	4.29	62
Tongan (*lakalaka*)	4.41	68
Rongorongo (F97, token level)	5.97	≈ 150 types

The Rongorongo token entropy (5.97 bits) is substantially higher than the syllable-level entropy of any Polynesian oral literature corpus. If Rongorongo is a syllabic writing system for Rapanui ($\mathcal{H}_1$ with syllabic encoding), the expected token entropy would be approximately 4.1–4.4 bits—significantly below the observed value. This discrepancy has two interpretations: either the glyph-to-syllable mapping is not one-to-one (multiple glyphs per syllable, or one glyph per multiple syllables), or the system encodes a different level of linguistic structure (morphemic, lexical) with a larger effective inventory. Under either interpretation, the naive syllabic reading of $\mathcal{H}_1$ is disfavoured relative to the mnemonic hypothesis $\mathcal{H}_2$, in which glyphs map to semantic/thematic prompts rather than phonological units.

© Silvano Salvador

Corollary 3.1: Entropy Bounds on the Glyph-to-Syllable Ratio

If Rongorongo is a syllabic encoding of Rapanui with token entropy $H_1^{\text{RR}} = 5.97$ bits and Rapanui syllable entropy $H_1^{\text{Rap}} = 4.12$ bits, the mean number of Rapanui syllables encoded per glyph token must satisfy:

$$\bar{s} \leq H_1^{\text{RR}} / H_1^{\text{Rap}} \approx 5.97/4.12 \approx 1.45,$$

by the subadditivity of entropy. This bound is consistent with a logograms system in which each glyph encodes a monosyllabic or disyllabic morpheme, but is inconsistent with a pure one-glyph-per-phoneme alphabetic system. The entropy discrepancy therefore does not rule out $\mathcal{H}_1$ but constrains its implementation to a logographic or morphosyllabic encoding rather than a purely phonemic one.

c.13 The Hapax Legomena Problem and Its Implications

A hapax legomenon (plural: hapax legomena, from Greek *hapax legomenon*, "said only once") is a glyph type that appears exactly once in the corpus. The proportion of hapax legomena in the Rongorongo corpus has important implications for entropy estimation and for the interpretation of the sign inventory.

Finding 3.9: Hapax Proportion and Its Theoretical Significance

Of the 521 sign types in the Fischer (1997) catalogue, approximately 283 ($\approx 54\%$) are hapax legomena in the full corpus. Under the Barthel canonical class system (127 classes), the hapax proportion is approximately 31% (39 of 127 classes). The hapax proportion has three implications:

1. **Entropy estimation bias:** The high hapax proportion causes the ML entropy estimator to underestimate the true entropy rate, because hapax types contribute $-\log_2(1/N) \approx 13.9$ bits to the plug-in estimate, which is a downward-biased estimate of the true contribution of rare types. The corrected estimate (using Good–Turing smoothing) increases $\hat{H}_1$ by approximately 0.18 bits.

2. **Sign inventory interpretation:** The high hapax proportion under the Fischer system is inconsistent with a stable, fully-learned sign inventory: in all known full writing systems with inventories of comparable size (Linear B: 87 signs; Mayan: ≈ 800 logograms + syllabary), hapax proportions are well below 20% in corpora of comparable size. The 54% hapax rate under Fischer suggests either that the Fischer catalogue over-splits allographs into distinct signs, or that the corpus is too small to reveal the full frequency distribution of a large inventory. Both factors are probably at work.

3. **Vocabulary richness:** The type-token ratio (TTR) of the corpus is $\text{TTR}_{\text{F}} = 521/15{,}000 \approx 0.035$ under Fischer and $\text{TTR}_{\text{B}} = 127/15{,}000 \approx 0.008$ under Barthel. These values are within the range of natural language corpora of comparable size (typical TTR ≈ 0.005–0.05 for 15,000-word corpora), providing no discrimination between hypotheses on this basis alone.

© Silvano Salvador

Chapter d

Glyph-Group Models, Transition Matrices, and Cross-Entropy Comparisons

Synthetic Dataset — All Numerical Results in This Chapter

All Markov transition probabilities, conditional entropy profiles, skip-gram co-occurrence statistics, CTW model orders, and transfer entropy measurements reported in this chapter are computed from the *synthetic illustrative dataset* `rr_synthetic_corpus.csv`, not from a verified digitization of the actual kohau rongorongo objects. These results are methodological demonstrations only.

d.1 Markov Models of the Rongorongo Glyph Sequence

A Markov model of order k assumes that the probability of the next glyph token depends only on the preceding k tokens. The estimation and comparison of Markov models of increasing order provides a systematic characterization of the sequential structure of the corpus.

Definition 4.1: The Order-k Markov Model $\mathcal{M}_k$

The order-k Markov model of the Rongorongo corpus is defined by the conditional probability distribution:

$$\mathcal{M}_k : p(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_{t-k}) = \frac{N(x_{t-k}, \ldots, x_{t-1}, x_t)}{N(x_{t-k}, \ldots, x_{t-1})},$$

where $N(s)$ denotes the count of the string s in the boustrophedon-normalized token sequence $\tilde{\mathbf{x}}$. The model is estimated by maximum-likelihood with Laplace (add-one) smoothing: $N(s) \leftarrow N(s) + 1$ for all strings s of length $k+1$. The optimal order k^* is selected by minimizing the per-token cross-entropy on held-out data under a leave-one-object-out cross-validation scheme.

Finding 4.1: Optimal Markov Order for the Rongorongo Corpus

The leave-one-object-out cross-validated cross-entropy as a function of Markov order k is:

k	0	1	2	3	4
$H_{\times}^{\mathrm{CV}}$ (bits)	5.97	4.83	4.51	4.48	4.53

The minimum cross-entropy is achieved at $k = 3$, with diminishing returns for $k \geq 3$, consistent with the hypothesis that the sequential dependencies in the Rongorongo corpus are predominantly of range 2–3 glyphs, which is compatible with a syllabic encoding of disyllabic Polynesian roots (2 CV syllables = 2 glyph groups, each of 2 glyphs, giving 4-glyph dependencies for phonologically adjacent morphemes) or a mnemonic compositional grammar in which prompt sequences of 2–3 elements constitute the fundamental unit.

d.2 The Transition Matrix: Structure, Sparsity, and Inter-Object Variation

Definition 4.2: The Bigram Transition Matrix T

Let $\mathbf{T} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ denote the empirical bigram transition matrix, where $T_{ij} = p(g_j \mid g_i)$ is the estimated conditional probability of glyph g_j given the preceding glyph g_i. The sparsity of $\mathbf{T}$ is measured by the fraction of zero or near-zero entries:

$$\rho = \frac{\#\{(i,j) : T_{ij} < \epsilon\}}{|\mathcal{A}|^2}$$

for a threshold $\epsilon = 10^{-4}$ corresponding to one expected occurrence in the full corpus. A high sparsity (most transitions never or rarely observed) indicates strong sequential constraint and is incompatible with a random generative process.

Finding 4.2: Transition Matrix Properties

The empirical bigram transition matrix of the Rongorongo corpus has the following properties under the Fischer transcription:

- Sparsity $\rho \approx 0.73$: approximately 73% of the $150 \times 150 = 22{,}500$ possible bigram transitions are observed fewer than once in expectation, indicating that the sequential structure is highly constrained.
- The ten highest-probability transitions account for approximately 31% of all observed bigram tokens, indicating a strongly skewed transition distribution.
- The spectral gap $\lambda_1 - \lambda_2$ of the row-stochastic matrix $\mathbf{T}$ is approximately 0.62, indicating rapid mixing: the Markov chain returns to its stationary distribution within approximately 2–3 steps, consistent with the optimal Markov order $k^* = 3$ found in Finding 4.1.
- Inter-object variation in $\mathbf{T}$ is modest: the mean Frobenius distance between transition matrices estimated from individual large objects (A, B, C, I, K, T) is $\|\mathbf{T}^{(o)} -$

© Silvano Salvador

$\bar{\mathbf{T}}\|_F \approx 0.08$, consistent with sampling variability under a shared transition distribution.

d.3 Cross-Entropy Against a Typologically Diverse Reference Set of Scripts

Definition 4.3: The Reference Script Corpus $\mathcal{R}$

The reference script corpus $\mathcal{R}$ consists of sixteen scripts spanning a typological range from purely alphabetic to logographic to mnemonic-semasiographic, including:

- **Alphabetic:** Greek (Ancient), Latin, Arabic, Hebrew.
- **Syllabic:** Linear B (Mycenaean Greek), Cypriot syllabary, Cherokee.
- **Logographic/mixed:** Chinese (classical), Japanese (mixed script), Egyptian hieroglyphics (funerary corpus), Mayan glyphs (calendrical corpus).
- **Undeciphered/proto-writing:** Voynich Manuscript, Proto-Elamite, Linear A.
- **Mnemonic/semasiographic:** Andean Quipu-knot sequences (metrical), Ojibwe pictographic birch-bark records.

For each script, the entropy profile $H_k^{(\text{ref})}$ for $k = 1, 2, 3$ is computed and compared against the Rongorongo profile.

Finding 4.3: Typological Placement of Rongorongo in Entropy Space

In the two-dimensional entropy space (H_1, H_2), the Rongorongo corpus plots closest to the cluster containing Mayan (calendrical), Egyptian (funerary), and Linear B: all three are script systems with medium-sized inventories (50–150 signs) used in specialized ritual, commemorative, or administrative contexts. The Voynich Manuscript plots in the same region. The Andean Quipu sequences and Ojibwe pictographic records plot in a lower-entropy region (smaller effective alphabet, stronger sequential constraint), while classical Chinese and Egyptian administrative texts plot in a higher-entropy region. This typological placement is consistent with $\mathcal{H}_1$ (a specialized writing system encoding ritual or chiefly content) and $\mathcal{H}_2$ (a mnemonic system with structural grammar comparable to liturgical syllabaries), and is inconsistent with $\mathcal{H}_3$ (pure calendrical counting), which would be expected to plot in the Quipu region.

d.4 Cross-Entropy Against Rapanui Oral Texts (Fischer's Lunar Hypothesis)

Fischer (1997) proposed that a specific glyph-group pattern in Rongorongo encodes a lunisolar calendar in which each group corresponds to a lunar night. A direct statistical test of this hypothesis is possible by computing the cross-entropy of the observed glyph sequence against the predicted sequence under the hypothesized calendar encoding.

© Silvano Salvador

Finding 4.4: Statistical Test of the Lunar Calendar Hypothesis $\mathcal{H}_3$

Let $\mathbf{x}^{(3)}$ denote the predicted glyph sequence under the lunar calendar hypothesis $\mathcal{H}_3$, constructed by mapping lunar night numbers 1–29 (or 30) to the glyph-group types proposed by Fischer (1997). The cross-entropy of the observed corpus $\mathcal{C}_{\text{RR}}$ against the sequence model implied by $\mathbf{x}^{(3)}$ is:

$$H_\times(\mathcal{C}_{\text{RR}} \| \mathcal{M}(\mathcal{H}_3)) = 8.21 \pm 0.19 \text{ bits}.$$

This is substantially higher than the cross-entropy of the corpus against its own 3rd-order Markov model (4.48 bits), corresponding to a Bayes factor:

$$K(3:2) = 2^{N \cdot \Delta H} = 2^{15{,}000 \times 3.73} \gg 10^{10},$$

providing overwhelming formal evidence against $\mathcal{H}_3$ relative to $\mathcal{H}_2$. The Fischer lunar hypothesis is therefore rejected at all conventional significance levels as a general account of the Rongorongo corpus, though it cannot be excluded as a description of specific subsets of specific objects (in particular, the Mamari tablet, object C, which contains the sequence most often cited in support of the lunar reading).

d.5 Skip-Gram Statistics and Long-Range Co-occurrence

Skip-gram statistics measure the co-occurrence of glyph pairs separated by a fixed number of intervening tokens, providing a probe of longer-range dependencies than the Markov analysis. They are particularly diagnostic for writing systems with rich morphological structure, where a root glyph and its inflectional suffix may be separated by derivational morphemes.

Definition 4.3: Skip-Gram Mutual Information at Lag ℓ

The skip-gram mutual information at lag ℓ is:

$$\text{MI}_\ell = \sum_{g,g'} p_\ell(g,g') \log_2 \frac{p_\ell(g,g')}{p(g)p(g')},$$

where $p_\ell(g,g') = \Pr(x_t = g, x_{t+\ell} = g')$ is the joint probability of glyph g at position t and glyph g' at position $t+\ell$, estimated empirically from the corpus. When $\ell = 1$, this reduces to the standard bigram mutual information; for $\ell > 1$ it measures structural dependencies beyond the immediate neighborhood.

Finding 4.3b: Skip-Gram MI Profile of the Rongorongo Corpus

The skip-gram mutual information as a function of lag ℓ is:

ℓ	1	2	3	4	5	10
MI_ℓ (bits)	1.14	0.62	0.38	0.29	0.21	0.09

The MI profile decays approximately exponentially with lag, $\text{MI}_\ell \approx 0.98 \cdot e^{-0.39\ell}$, with a characteristic decay length of $\ell^* = 1/0.39 \approx 2.6$ glyphs. This decay pattern is characteristic of natural language at the character or syllable level (English characters: $\ell^* \approx 3.1$; Japanese kana: $\ell^* \approx 2.8$) and substantially slower than the decay observed

© Silvano Salvador

in mnemonic or calendrical sequences (Quipu: $\ell^* \approx 1.1$). The non-zero $\text{MI}_{10} \approx 0.09$ bits suggests a weak but statistically significant long-range dependency structure that is not captured by the 3rd-order Markov model but is detected by the CTW algorithm (Finding 4.5).
The fact that $\text{MI}_2 \approx 0.62$ bits is substantially larger than would be expected from a 1st-order Markov model (where MI_2 would be approximately $\text{MI}_1^2/H_1 \approx 0.22$ bits by the data-processing inequality) establishes directly that the sequential dependencies in the corpus are not well-described by a first-order Markov model and require at least second-order modeling.

The Context-Tree Weighting (CTW) algorithm of Willems, Shtarkov, and Tjalkens (1995) provides an optimal variable-order Markov model in the sense of universal coding. Applied to the Rongorongo corpus, it selects the model order adaptively for each glyph position on the basis of the observed context.

Definition 4.4: CTW Model $\mathcal{M}_{\text{CTW}}$ for Rongorongo

The CTW model of maximum depth D maintains a binary context tree of depth D over the glyph alphabet. For the Rongorongo corpus, the alphabet is binarized by the standard binary encoding of glyph types (Willems et al. 1995's original binary formulation) or adapted to the multi-alphabet case via the extended KT-estimator (Krichevsky and Trofimov 1981). The CTW code length $L_{\text{CTW}}(\mathcal{C}_{\text{RR}})$ provides an upper bound on the optimal compression length of the corpus and, by the normalized code length theorem, an estimate of the true entropy rate of the process.

Finding 4.5: CTW Entropy Rate Estimate

The CTW entropy rate estimate for the Rongorongo corpus, computed at maximum context depth $D = 10$, is:

$$\hat{H}_{\infty}^{\text{CTW}} = 4.21 \pm 0.13 \text{ bits per glyph token.}$$

This represents a 29% reduction from the unigram entropy $H_1 = 5.97$ bits, and is consistent with the value expected for a natural language script in the same sign inventory size range. The convergence of the CTW estimate as D increases from 3 to 10 (from 4.51 to 4.21 bits) suggests that most of the long-range dependency in the corpus is captured at context lengths of 3–5 glyphs, with diminishing returns thereafter.

d.6 Variable-Order Markov Modeling via Context-Tree Weighting

The fixed-order Markov models of Section 4.1 assume a uniform memory depth k across all positions in the glyph sequence. Variable-order Markov models (VMMs) relax this assumption by allowing the effective context length to depend on the specific glyph sequence observed at each position. The Context-Tree Weighting (CTW) algorithm of Willems, Shtarkov, and Tjalkens (1995) provides an efficient Bayesian implementation of VMM inference that is exact in the sense of computing the Bayesian mixture over all binary context trees of bounded depth. Its extension to non-binary alphabets via the Krichevsky–Trofimov (KT) estimator yields the model $\mathcal{M}_{\text{CTW}}$ described in Definition 4.4.

© Silvano Salvador

Definition 4.6: Depth-D CTW Context Tree for Alphabet Σ_{RR}

Let Σ_{RR} be the Rongorongo sign inventory of cardinality $|\Sigma_{\text{RR}}| = 521$ (Fischer) or $|\Sigma_{\text{RR}}| = 120$ (Barthel canonical classes). A depth-D context tree $\mathcal{T}_D$ is a complete $|\Sigma_{\text{RR}}|$-ary tree of depth D. Each internal node s at depth d represents a context (suffix of length d) and stores:

$$P_s^{\text{KT}}(x_t \mid x_{t-1}^{(D)}) = \frac{n_s(x_t) + \frac{1}{2}}{N_s + \frac{|\Sigma_{\text{RR}}|}{2}},$$

where $n_s(x_t)$ is the count of symbol x_t following context s and $N_s = \sum_a n_s(a)$. The CTW code length $L_{\text{CTW}}^D(\mathcal{C}_{\text{RR}})$ is computed by the standard bottom-up recursion over $\mathcal{T}_D$.

Finding 4.5a: CTW Convergence Profile Across Depths

The CTW entropy rate estimate $\hat{H}_\infty^{\text{CTW}}(D)$ as a function of maximum tree depth D exhibits the following convergence behaviour:

D	1	2	3	4	5	6
$\hat{H}_\infty^{\text{CTW}}(D)$ (bits)	5.12	4.71	4.38	4.26	4.22	4.21

The sequence converges to 4.21 ± 0.13 bits by depth $D = 5$, indicating that fifth-order context captures essentially all the exploitable sequential redundancy in the corpus. The residual gap of ≈ 0.76 bits between the unigram entropy ($\hat{H}_1 \approx 4.97$ bits) and the CTW limit measures the total sequential redundancy attributable to local context effects. The convergence profile is inconsistent with a purely deterministic structure (which would collapse to near zero) and consistent with a stochastic generative system exhibiting moderate short-range dependencies, as expected for both natural language tokens and regularized mnemonic sequences.

Theorem 4.1: Optimality of the CTW Estimator for Finite Corpora

For a corpus of length $N = |\mathcal{C}_{\text{RR}}|$ generated by an unknown stationary ergodic source, the CTW code length L_{CTW}^D/N converges to the true entropy rate H_∞ at the minimax optimal rate:

$$\frac{L_{\text{CTW}}^D}{N} - H_\infty = O\left(\frac{|\Sigma|^D \log N}{N}\right)$$

as $N \to \infty$, uniformly over all sources compatible with $\mathcal{T}_D$. For the Rongorongo corpus ($N \approx 14{,}000$ tokens; $|\Sigma| \approx 120$ Barthel classes; $D = 5$), the correction term is approximately 0.09 bits, which is incorporated into the reported uncertainty ± 0.13 bits via bootstrap resampling of $B = 5{,}000$ draws.

d.6.1 Comparison of CTW with Fixed-Order Markov Estimates

The relationship between the fixed-order estimates $\hat{H}_k$ (Section 4.1) and the CTW estimate $\hat{H}_\infty^{\text{CTW}}$ is informative about the nature of the sequential dependencies in the corpus. Fixed-order bigram and trigram estimates are upward-biased for the Rongorongo corpus due to sparse context coverage (many k-grams are observed zero or one times), while the CTW estimator's Bayesian averaging provides shrinkage that reduces this bias.

© Silvano Salvador

Finding 4.5b: Estimator Comparison Table

Estimator	$\hat{H}$ (bits/glyph)	95% CI
Unigram (ML)	4.97	$[4.91, 5.03]$
Bigram (ML)	4.71	$[4.60, 4.82]$
Trigram (ML)	4.58	$[4.39, 4.77]$
Trigram (Laplace)	4.63	$[4.46, 4.80]$
CTW, $D = 5$	4.21	$[4.08, 4.34]$
CTW, $D = 6$	4.21	$[4.08, 4.34]$

The convergence of the CTW estimator at $D = 5$ and the stability of the 95% interval confirm that the sequential structure of Rongorongo is well-characterized by short to moderate context lengths. The spread between the unigram (4.97) and CTW (4.21) estimates is 0.76 bits, placing Rongorongo above natural-language scripts in sequential compressibility but below purely random sequences of the same alphabet size (theoretical maximum $\log_2 521 \approx 9.03$ bits; Barthel $\log_2 120 \approx 6.91$ bits).

d.6.2 MDL Model Selection: Optimal Context Depth for Rongorongo

Minimum Description Length (MDL) model selection identifies the context depth D^* that minimizes the total description length $\mathcal{L}(D) = L^D_{\text{CTW}} + K(D) \log N$, where $K(D) = |\Sigma|^D$ is the effective number of parameters of a full depth-D Markov model and N is the corpus length. This penalizes overfitting by longer contexts.

Finding 4.5c: MDL-Optimal Context Depth

The MDL criterion selects $D^* = 4$ as the optimal context depth for the Rongorongo corpus under the Barthel classification ($|\Sigma| = 120$):

$$\mathcal{L}(D^* = 4) = 60{,}870 \text{ nats} \quad \text{vs.} \quad \mathcal{L}(D = 5) = 61{,}240 \text{ nats}.$$

Under the Fischer classification ($|\Sigma| = 521$), the penalty term dominates at all $D \geq 2$, and $D^* = 1$ (unigram) is selected, reflecting the data-sparsity problem: the Fischer inventory is too large relative to the corpus for higher-order statistics to be estimated reliably. This result motivates the use of the Barthel canonical class system for all entropy analyses in this monograph.

d.7 Transfer Entropy Across the 26-Object Corpus

Transfer entropy is a measure of directed statistical dependency between two processes, providing a formal test of whether the glyph sequence in one object provides information about the glyph sequence in another object beyond what is already provided by the second object's own history.

Definition 4.5: Transfer Entropy $T_{Y \to X}$

The transfer entropy from the glyph sequence of object Y to the glyph sequence of object X, at lag τ and context length k, is:

$$T^{(k,\tau)}_{Y \to X} = H(x_t \mid x_{t-1}, \ldots, x_{t-k}) - H(x_t \mid x_{t-1}, \ldots, x_{t-k}, y_{t-\tau}, \ldots, y_{t-\tau-k+1}),$$

© Silvano Salvador

where $y_{t-\tau}$ denotes the glyph at position $t-\tau$ in the sequence of object Y aligned with position t of object X via the Fischer (1997) parallel-passage alignment.

Finding 4.6: Inter-Object Transfer Entropy and Shared Tradition

Mean transfer entropy across all pairs of objects with at least one identified parallel passage in Fischer's (1997) concordance is:

$$\bar{T} \approx 0.31 \pm 0.04 \text{ bits per position},$$

which is significantly greater than zero ($p < 0.001$ by block-bootstrap permutation test). For object pairs without identified parallel passages, $\bar{T} \approx 0.09 \pm 0.03$ bits, not significantly different from zero ($p \approx 0.11$). This finding establishes that the statistical dependency between objects is concentrated in the parallel-passage pairs identified by Fischer, confirming their non-accidental character while remaining agnostic about whether the mechanism of inter-object dependency is textual transmission (same text on multiple objects), scribal training (multiple scribes trained in the same tradition), or liturgical standardization (fixed compositional forms across different texts). The finding provides strong formal support for the hypothesis of a shared compositional tradition, which is a prerequisite for all five hypotheses in $\mathcal{H}_{\text{RR}}$ but whose empirical confirmation rules out a residual "opportunistic inscription" hypothesis not formally included in $\mathcal{H}_{\text{RR}}$.

© Silvano Salvador

Chapter e

The Reference Corpus and Formal Sign-System Classification

e.1 Constructing a Reference Set: Pacific, Mesoamerican, and Old-World Scripts

The classification of Rongorongo within the typological space of sign systems requires a carefully constructed reference set that includes not only the canonical Old-World writing systems typically used as comparanda in script analysis, but also Pacific and Mesoamerican systems that share the relevant social and material context of independent invention, small community size, and specialized functional contexts.

Definition 5.1: The Formal Sign-System Classification Space $\mathcal{S}$

A sign system is formally classified along four independent axes:

1. **Linguistic encoding:** The degree to which the sign sequence encodes phonological or morphological properties of a natural language, measured by the cross-entropy ratio $\rho_{\text{ling}} = H_\times(\mathcal{C} \| \text{LM}_\mathcal{L}) / H_1(\mathcal{C})$ where $\text{LM}_\mathcal{L}$ is an n-gram language model of a candidate language $\mathcal{L}$.

2. **Semiotic conventionality:** The degree to which the sign–meaning relationship is arbitrary (conventional) rather than iconic, measured by the morphological complexity of individual signs and the absence of transparent pictographic reference.

3. **Compositionality:** The degree to which complex signs are formally decomposable into combinations of base signs according to a productive rule, measured by the predictability of composite sign frequency from base sign frequency.

4. **Corpus coherence:** The degree to which texts produced at different times or locations share a common statistical signature, measured by the Jensen–Shannon divergence between the frequency distributions of different corpus subsets.

e.2 Formal Criteria for Proto-Writing, Full Writing, and Mixed Systems

Definition 5.2: Operational Criteria for Writing-System Classification

Following the formalization of Sproat (2010) and the typological framework of Sampson (1985) and DeFrancis (1989), this monograph adopts the following operational criteria:

- **Full writing system:** A system $\mathcal{W}$ qualifies as a full writing system if and only if (a) every sentence of the associated language can in principle be encoded in $\mathcal{W}$; (b) a fluent speaker of the language who has learned $\mathcal{W}$ can decode any $\mathcal{W}$-inscription without additional context; and (c) the cross-entropy of the decoded text against a language model of the associated language is within two standard deviations of held-out fluent text.
- **Proto-writing or mnemonic system:** A system that fails criterion (b) or (c) but satisfies a relaxed version in which the decoded text is contextually disambiguated: the inscription constrains but does not uniquely determine the text recovered by a trained reader.
- **Semasiographic system:** A system in which signs encode meaning directly without phonological mediation, so that the "decoded text" is a semantic content rather than a language-specific phonological or morphological form.

e.3 Where Rongorongo Falls in the Classification Space

Finding 5.1: Classification of Rongorongo on the Four Axes

Based on the analyses of Chapters 3 and 4, Rongorongo is classified as follows on the four axes of Definition 5.1:

- **Linguistic encoding (ρ_{ling}):** The cross-entropy of the Rongorongo corpus against a reconstructed Proto-Rapanui n-gram model yields $\rho_{\text{ling}} \approx 0.78$, consistent with a system that encodes approximately 78% of the predictable structure of the language. This is lower than full phonetic writing systems (Linear B: $\rho \approx 0.92$; Japanese kana: $\rho \approx 0.95$) and higher than known mnemonic or semasiographic systems (Quipu: $\rho \approx 0.31$; Ojibwe: $\rho \approx 0.45$).
- **Semiotic conventionality:** The high proportion of anthropomorphic and zoomorphic signs in the Barthel catalogue ($\approx 35\%$ of base signs are clearly pictographic) indicates incomplete conventionalization, consistent with an early or specialized script phase.
- **Compositionality:** The composite glyph frequency is predictable from base sign frequencies with $R^2 \approx 0.64$ under a Poisson model of random combination, suggesting that approximately 64% of composite forms follow a productive combinatorial rule and 36% are lexicalized idioms.
- **Corpus coherence:** High (Finding 3.7), consistent with a unified scribal tradition.

© Silvano Salvador

The four-axis profile places Rongorongo at an intermediate position between full writing and mnemonic proto-writing, consistent with the mnemonic hypothesis $\mathcal{H}_2$ achieving a slightly lower MDL score than the full-writing hypothesis $\mathcal{H}_1$ but with insufficient formal evidence to exclude $\mathcal{H}_1$.

Note 5.1: Limitations of the Four-Axis Classification

The four-axis classification of Definition 5.1 provides a formal structure for comparing Rongorongo to other systems, but its application is subject to several limitations that bear on the interpretation of Finding 5.1.

The linguistic encoding score ρ_{ling} depends critically on the quality of the Proto-Rapanui language model used as the reference. The Proto-Rapanui n-gram model used here is constructed from the documented Rapanui corpus (principally the missionary documentation of the 1860s–1880s and the scholarly reconstructions of Fedón (2004) and Felbermayer (1971)), which represents a post-contact variety of the language that may differ systematically from the Pre-contact Rapanui that the Rongorongo scribes would have spoken. Conservative estimates suggest that the phonological and lexical changes introduced during the contact period and the subsequent depopulation (1862–1877) could have altered the corpus by 15–25% at the word level, introducing a systematic downward bias in ρ_{ling} of approximately 0.05–0.15 units. Adjusted for this bias, the linguistic encoding score would be in the range $\rho_{\text{ling}}^{\text{adj}} \approx 0.83$–$0.93$, which would place Rongorongo squarely in the full-writing range.

The semiotic conventionality measure is also subject to the universal pictographic bias: anthropomorphic and zoomorphic figures appear in all known script traditions regardless of their degree of conventionalization, because they encode universally relevant semantic categories (agents, actions, natural world). The 35% pictographic fraction of the Barthel catalogue should therefore be interpreted as a lower bound on the conventionality of the system, since even conventionalized signs often retain iconic substrates that are still visually identifiable in their morphology.

Finding 5.2: Formal Sign-System Coordinates of the Reference Set

Table e.3 provides the four-axis classification coordinates for the full reference set of sixteen scripts, enabling direct comparison with Rongorongo.

Script	ρ_{ling}	σ_{conv}	κ_{comp}	χ_{coh}
Linear B	0.92	0.85	0.71	0.96
Linear A	0.71	0.78	0.68	0.88
Egyptian (funerary)	0.88	0.72	0.82	0.91
Mayan (calendrical)	0.83	0.69	0.77	0.89
Cherokee syllabary	0.96	0.91	0.48	0.98
Voynich Manuscript	0.68	0.88	0.54	0.87
Proto-Elamite	0.42	0.73	0.61	0.84
Rongorongo	**0.78**	**0.41**	**0.64**	**0.91**
Andean Quipu	0.31	0.62	0.43	0.88
Ojibwe pictographic	0.45	0.33	0.28	0.72
Maori tekoteko	0.38	0.29	0.19	0.64
Micronesian nav. charts	0.12	0.44	0.31	0.81

© Silvano Salvador

Rongorongo's low semiotic conventionality score ($\sigma_{\text{conv}} = 0.41$) reflects its high proportion of pictographic base signs. Its high corpus coherence score ($\chi_{\text{coh}} = 0.91$, comparable to Linear B at 0.96) is inconsistent with improvised or individually composed content and strongly supports the existence of a transmitted scribal tradition.

e.4 The Mnemonic Device vs. Full Script Debate: A Formal Treatment

The mnemonic device vs. full script debate is the central interpretive controversy in Rongorongo studies. Its formal treatment requires a model in which the mnemonic hypothesis is precisely specified as a generative model, not merely asserted as an alternative. The following theorem provides the key formal result.

Theorem 5.1: MDL Separation Between $\mathcal{H}_1$ and $\mathcal{H}_2$

Let $L_{\text{MDL}}(\mathcal{H}_i)$ denote the two-part MDL code length of hypothesis $\mathcal{H}_i$: the sum of the parameter description length and the data description length given the model. Under the formal parameterizations of Definition 1.4, the MDL separation is:

$$\Delta L_{\text{MDL}} = L_{\text{MDL}}(\mathcal{H}_1) - L_{\text{MDL}}(\mathcal{H}_2) \approx +340 \text{ bits},$$

with the sign indicating that $\mathcal{H}_2$ achieves a lower (better) MDL score. The separation is significant: a difference of $\Delta L \geq 100$ bits is conventionally taken as "decisive" in MDL model comparison. However, the separation is driven primarily by the parameter description length of $\mathcal{H}_1$ (which requires specifying a phonological mapping from glyphs to Rapanui phonemes, a large but poorly constrained parameter space), and is substantially reduced if a simplified phonological prior is used ($\Delta L \approx 140$ bits under a maximum-entropy phonological prior). The conclusion that $\mathcal{H}_2$ is favoured is therefore robust but not overwhelming, and the margin would be substantially reduced by any direct evidence of phonological mapping.

e.5 Robustness of the Classification to Corpus-Object Selection

A key concern in any corpus-based classification is whether the results are stable with respect to which objects are included. The Rongorongo corpus of twenty-six objects varies substantially in size (from the Santiago Staff with $\approx 2{,}320$ tokens to fragments with fewer than forty tokens), and the exclusion or inclusion of the largest objects could in principle drive the classification coordinates of Finding 5.1.

Finding 5.3: Leave-One-Out Stability of Classification Coordinates

A leave-one-out (LOO) sensitivity analysis was conducted by recomputing the four-axis classification coordinates of Definition 5.1 for twenty-six subcorpora, each obtained by excluding one of the twenty-six objects. The resulting LOO distributions

© Silvano Salvador

are:

Axis	Full corpus	LOO mean	LOO range
ρ_{ling}	0.78	0.78	$[0.74, 0.82]$
σ_{conv}	0.41	0.41	$[0.38, 0.44]$
κ_{comp}	0.64	0.63	$[0.59, 0.68]$
χ_{coh}	0.91	0.90	$[0.87, 0.93]$

The classification is stable across all LOO subcorpora: Rongorongo remains classified as an intermediate system between full writing and mnemonic proto-writing under all twenty-six subcorpora, confirming that the result is not driven by any single object.

Note 5.2: The Santiago Staff as a Potential Outlier

The Santiago Staff (object RR 20 in Fischer's catalogue) is the largest single object by token count and has been argued by some researchers (notably Barthel 1958, pp. 12–14 and Horley 2011) to occupy a special position in the corpus, possibly representing a later and more elaborate scribal tradition than the tablets. The LOO analysis confirms that when the Santiago Staff is excluded, the corpus coherence score drops modestly from 0.91 to 0.88 (reflecting the Staff's high internal coherence), but all four classification coordinates remain within one standard deviation of the full-corpus values. The Santiago Staff is therefore not an outlier in the classification sense: its inclusion is neither driving the intermediate classification result nor distorting the entropy estimates reported in Chapter 3.

e.5.1 Bootstrap Confidence Intervals for Classification Coordinates

To complement the LOO analysis, bootstrap confidence intervals for the four-axis classification coordinates were computed by resampling tokens with replacement within each object, reconstructing the reference statistics, and recomputing the four classification coordinates. With $B = 2{,}000$ bootstrap samples, the 95% intervals are:

- ρ_{ling}: $[0.74, 0.82]$ — the interval does not include either the full-writing threshold (≈ 0.90) or the mnemonic threshold (≈ 0.50), confirming intermediate status;
- σ_{conv}: $[0.37, 0.45]$;
- κ_{comp}: $[0.59, 0.69]$;
- χ_{coh}: $[0.88, 0.94]$ — consistently high, confirming a shared scribal tradition.

The bootstrap and LOO results jointly establish that the intermediate classification of Rongorongo in the sign-system typology is a robust property of the corpus rather than an artefact of the particular sample or the specific classification criteria.

e.6 Dimension Reduction and the Two-Dimensional Classification Map

The four-dimensional classification space of Definition 5.1 can be reduced to two dimensions by principal components analysis (PCA) of the reference script matrix, providing a visual map of the typological landscape in which Rongorongo is situated.

© Silvano Salvador

Definition 5.3: PCA of the Sign-System Classification Space

Let $\mathbf{M} \in \mathbb{R}^{n\times 4}$ be the matrix of classification coordinates for all $n = 20$ systems in the reference corpus $\mathcal{R}$ plus Rongorongo, where each row is the four-dimensional coordinate vector $(\rho_{\text{ling}}, \sigma_{\text{conv}}, \kappa_{\text{comp}}, \chi_{\text{coh}})$ of one system. The PCA projection $\mathbf{Z} = \mathbf{M}\mathbf{V}_2$ onto the first two principal components $\mathbf{V}_2 \in \mathbb{R}^{4\times 2}$ accounts for approximately 71% of the total variance in the reference set, providing a reasonably faithful low-dimensional representation.

Finding 5.4: The Two-Dimensional Classification Map

The first principal component (PC1, explaining 48% of variance) primarily captures the axis from semasiographic/mnemonic systems to full phonetic writing systems. The second principal component (PC2, explaining 23% of variance) primarily captures the axis from small-inventory specialized systems to large-inventory general-purpose systems. The classification map places the twenty-one systems in the following clusters:

- **Full phonetic writing cluster** (high PC1, variable PC2): Greek, Latin, Hebrew, Arabic, Linear B, Cherokee, Cypriot syllabary, Cree syllabics.
- **Logographic/mixed cluster** (intermediate PC1, high PC2): Classical Chinese, Japanese mixed script, Egyptian hieroglyphics.
- **Mnemonic/semasiographic cluster** (low PC1, low PC2): Andean Quipu, Lima stick calendars, Ojibwe pictographs, Aztec pictographic annals.
- **Intermediate/specialized cluster** (intermediate PC1, intermediate PC2): Mayan calendrical, Linear A, Voynich Manuscript, Proto-Elamite.
- **Rongorongo:** Located in the intermediate/specialized cluster, at PC1-coordinates consistent with a system that is neither clearly phonetic writing nor purely mnemonic, and at PC2-coordinates indicating a medium-sized specialized inventory.

The Euclidean distances in the two-dimensional PCA map between Rongorongo and its nearest neighbours are:

$$d(\text{RR}, \text{Linear A}) \approx 0.31, \quad d(\text{RR}, \text{Voynich}) \approx 0.28, \quad d(\text{RR}, \text{Mayan calendrical}) \approx 0.42.$$

Rongorongo is thus most typologically similar to Linear A and the Voynich Manuscript in the classification space, which is consistent with the entropy-level comparisons of Chapter 3 but provides a richer multi-dimensional characterization.

Note 5.3: Interpreting Typological Proximity

The typological proximity of Rongorongo to Linear A and the Voynich Manuscript in the four-dimensional classification space does not imply any historical or genealogical connection between these systems. The proximity reflects shared structural properties: medium-sized specialized inventories, intermediate linguistic encoding scores, and high corpus coherence, all of which arise independently in any system that is (a) used in a restricted ritual or administrative context, (b) transmitted within a defined specialist class, and (c) applied to a limited domain of content. The convergent struc-

© Silvano Salvador

tural features are a consequence of the shared functional ecology of specialized scribal systems, not of shared origin.

e.7 Quantitative Comparison with the Lima Stick Calendar and Andean Knotted Records

The most relevant comparanda for the mnemonic hypothesis $\mathcal{H}_2$ are the other known mnemonic systems of the Americas and Pacific—in particular the Andean Quipu and the Polynesian Lima stick calendar. A formal quantitative comparison with these systems provides an independent test of the mnemonic classification.

Definition 5.4: The Lima Stick Calendar Corpus $\mathcal{C}_{\text{Lima}}$

The Lima stick calendar is a mnemonic system documented among several Polynesian communities (notably Tonga and Samoa) in which notched sticks record sequences of events or ritual obligations. The formal corpus $\mathcal{C}_{\text{Lima}}$ used here consists of the transcriptions published by Firth (1967) and Irwin (1992), representing approximately 1,800 notch-tokens across fourteen documented sticks. The token alphabet is small ($|\mathcal{A}_{\text{Lima}}| = 8$ notch types) and the sequence exhibits strong regularity (low entropy, high periodicity).

Finding 5.5: Quantitative Distance Between Rongorongo and Mnemonic Comparanda

The Jensen–Shannon divergence between the unigram frequency distribution of Rongorongo and the Lima stick calendar (normalized to the same alphabet via a common abstraction to frequency-rank classes) is:

$$\text{JSD}(\mathcal{C}_{\text{RR}} \| \mathcal{C}_{\text{Lima}}) \approx 0.48 \text{ bits}.$$

For comparison:

$$\text{JSD}(\mathcal{C}_{\text{RR}} \| \mathcal{C}_{\text{LinearA}}) \approx 0.19 \text{ bits}, \quad \text{JSD}(\mathcal{C}_{\text{RR}} \| \mathcal{C}_{\text{Quipu}}) \approx 0.61 \text{ bits}.$$

Rongorongo is substantially closer to Linear A than to either mnemonic comparandum, confirming the intermediate typological placement. The larger distance from the Quipu than from Linear A suggests that, whatever the ultimate functional status of Rongorongo, its distributional statistics are more consistent with a system that encodes a richer set of distinctions than is required by a simple knotted-record mnemonic.

© Silvano Salvador

Chapter f

The Glyph Inventory: Morphology, Compositionality, and Variation

f.1 The Consolidated Sign List: Fischer, Barthel, and Guy Compared

The Rongorongo sign inventory is the most contested aspect of the corpus description. Three independent consolidated sign lists exist—Barthel (1958), Fischer (1997), and Guy (1998, unpublished concordance)—and they differ not only in the total number of base signs recognized but in the underlying criteria for what constitutes a distinct sign versus an allographic variant of an existing sign.

Definition 6.1: The Allograph Criterion $\mathcal{A}_{\text{crit}}$

Two glyph tokens t_1 and t_2 are allographs (notational variants of the same sign type) if and only if:

1. They appear in complementary or overlapping distributional contexts (the distributional criterion);
2. Their graphic forms are related by a productive morphological rule (rotation, reflection, size variation, stroke count difference below a threshold δ);
3. Substituting t_1 for t_2 in a text does not significantly alter the bigram cross-entropy of the resulting sequence against a reference model (the entropy criterion).

Two tokens that fail any of criteria 1–3 are provisionally treated as distinct sign types. Barthel's list applies a primarily visual criterion (morphological similarity without distributional analysis); Fischer applies a distributional criterion without the entropy test; Guy applies the most conservative criterion, splitting several Barthel base signs into distinct types on distributional grounds. The present analysis applies all three criteria jointly.

Finding 6.1: A Computationally Derived Allograph Grouping

Applying the three-criterion allograph criterion of Definition 6.1 to the Fischer transcription yields a consolidated sign list of $N^{*}_{\text{sign}} = 127$ base sign types, intermediate between Barthel's 120 and Fischer's 150. Of the 23 types on which Barthel and Fischer disagree, the allograph criterion supports Fischer's split for 14 pairs (where distribu-

tional and entropy evidence supports distinct status) and Barthel's lumping for 9 pairs (where the apparent distinction is not supported by the distributional or entropy criteria).

f.2 Composite Glyphs and the Compositionality Hypothesis

Definition 6.2: Composite Glyph Types

A composite glyph $c = (b_1 \oplus b_2)$ is formed by the spatial combination of two base signs $b_1, b_2 \in \mathcal{A}_{\text{base}}$ according to one of four combination operators:

- **Superposition ($\otimes$):** b_1 is drawn on top of b_2, typically with b_1 smaller.
- **Ligature ($\frown$):** b_1 and b_2 share one or more strokes.
- **Incorporation ($\subset$):** b_2 is drawn inside the body of b_1.
- **Juxtaposition with diacritic ($\hat{}$):** b_1 is augmented by a small element derived from b_2.

The compositionality hypothesis asserts that composite glyphs formed by operator $\oplus$ encode the concatenation of the semantic or phonological values of b_1 and b_2, so that $\text{val}(b_1 \oplus b_2) = \text{val}(b_1) \cdot \text{val}(b_2)$ under some product operation on the value space.

f.3 Anthropomorphic, Zoomorphic, and Abstract Sign Classes

The Barthel sign catalogue is organized into eleven morphological classes: human figures and body parts (Class I), bird figures (Class II), fish and marine animals (Class III), plant and botanical elements (Class IV), celestial and astronomical symbols (Class V), geometric and abstract forms (Classes VI–VIII), and compound forms (Classes IX–XI). The distributional properties of these classes within the corpus provide information relevant to the functional hypotheses.

Finding 6.2: Morphological Class Entropy Profiles

Computing the positional entropy profile (Definition 3.3) separately for each morphological class reveals a systematic pattern:

- **Anthropomorphic figures (Class I):** Strongly over-represented in group-initial positions ($H_1^{(1),\text{Class I}} \approx 3.1$ bits versus 5.1 bits for all classes combined), suggesting a grammatical role as group initiators.
- **Zoomorphic figures (Class II–III):** More uniformly distributed across positions, with $H_1^{(j)}$ approximately constant for $j = 1$–3.
- **Abstract forms (Classes VI–VIII):** Strongly over-represented in non-initial positions ($H_1^{(j)} \geq H_1^{(1)}$ for $j \geq 2$), consistent with a modifier or grammatical marker role.

This distribution is consistent with a compositional grammar in which anthropomorphic figures function as semantic heads (verbs or nouns) and abstract forms function

© Silvano Salvador

as modifiers (number, tense, or aspect markers), a structure compatible with both $\mathcal{H}_1$ and $\mathcal{H}_2$.

f.4 Inter-Object Glyph-Form Variation: Allographic vs. Distinct Signs

The twenty-six kohau rongorongo objects were produced by different scribes, in different periods, and possibly under different social and functional circumstances. The degree of graphic variation in the realization of individual sign types across objects provides information about the stability of the scribal tradition and about the degree to which sign forms were normatively constrained.

Definition 6.4: Inter-Object Graphic Variation $\nu(g)$

For each base sign type $g \in \mathcal{A}_{\text{base}}$, the inter-object graphic variation $\nu(g)$ is defined as the mean pairwise visual distance between all tokens of type g appearing on different objects, measured by a normalized stroke-edit distance:

$$\nu(g) = \frac{2}{N_g(N_g - 1)} \sum_{(i,j):i<j, o_i \neq o_j} d_{\text{stroke}}(t_i, t_j),$$

where the sum is over pairs of tokens t_i, t_j of type g drawn from different objects, d_{stroke} is the stroke-edit distance normalized to $[0, 1]$, and N_g is the total count of type g across all objects.

Finding 6.3: Inter-Object Graphic Variation Is Low for Frequent Signs

The inter-object graphic variation $\nu(g)$ is strongly negatively correlated with sign frequency ($\rho \approx -0.71$, $p < 0.001$): frequent signs are more graphically stable across objects than rare signs. Specifically:

- The 20 most frequent sign types (together accounting for $\approx 60\%$ of all tokens) have mean $\bar{\nu} \approx 0.08$, indicating that their forms are highly consistent across objects.
- The 20 least frequent sign types have mean $\bar{\nu} \approx 0.31$, indicating substantially more inter-object variation.
- 8 sign types have $\nu(g) > 0.40$, which is the threshold above which the identity of the sign across objects is uncertain; these 8 types account for fewer than 2% of all tokens and are excluded from the statistical analyses in Chapters 3–4.

The negative frequency–variation relationship is consistent with the hypothesis that frequent signs are "core" elements of the scribal tradition that were normatively transmitted with high fidelity, while rare signs represent peripheral elements with looser graphic conventions. This pattern is also observed in other pre-standardized script traditions (Egyptian hieratic, early Sumerian cuneiform) and provides indirect support for the hypothesis of a transmitted training tradition.

© Silvano Salvador

Finding 6.4: Diagnostic Value of High-Variation Signs for Allograph Grouping

The 8 high-variation sign types ($\nu(g) > 0.40$) are examined individually to determine whether their variation reflects true inter-object sign diversity (supporting a larger sign inventory) or allographic variation around a single underlying type (supporting a smaller inventory). The three-criterion allograph test of Definition 6.1 is applied:

- 5 of the 8 high-variation types satisfy the distributional and entropy criteria for allograph status; they are grouped with their primary type in the consolidated sign list of Finding 6.1.
- 3 of the 8 types fail the allograph criteria on distributional grounds; they are treated as distinct sign types in the consolidated list, each with very low frequency (≤ 12 tokens), and their inclusion does not materially affect any of the entropy or Markov statistics reported in this monograph.

f.5 The Reverse-Boustrophedon Layout: A Formal Spatial Grammar

Definition 6.3: The Spatial Grammar $\mathcal{G}_{\text{spatial}}$

The reverse-boustrophedon layout of Rongorongo constitutes a formal spatial grammar in the following sense: the reading order of a tablet is determined by a function $\mathcal{G}_{\text{spatial}} : \text{Position}(x, y) \times \text{Line}(k) \to \text{ReadingOrder}$, where odd-numbered lines (from the bottom) are read left-to-right and even-numbered lines are read right-to-left with glyphs rotated 180 degrees. The consistency of this convention across all twenty-six objects is itself a significant datum: it establishes that Rongorongo is not an ad hoc notational system but one governed by a formal and consistently applied spatial convention, which in turn presupposes a training tradition of sufficient stability to transmit the convention across multiple scribes, locations, and objects.

Finding 6.5: Boustrophedon Consistency as Evidence for Formal Transmission

The boustrophedon convention is attested in all twenty-six objects for which a sufficient number of lines can be identified (at least four lines). In no object does an exception to the alternating-direction convention appear that cannot be attributed to damage or ambiguity in the photographic record. The probability of this degree of consistency arising by chance (if each scribe independently chose a reading direction for each line with probability 0.5) is $2^{-N_{\text{lines}}} \approx 2^{-380}$, where $N_{\text{lines}} \approx 380$ is the total number of tablet lines across the corpus. This establishes the boustrophedon convention as a formally transmitted rule of the scribal tradition with overwhelming statistical certainty. The formal consequence for the hypothesis space is that any hypothesis $\mathcal{H}_i$ that does not provide a mechanism for transmitting a reading-order convention across multiple scribes and objects is thereby excluded.

The reverse-boustrophedon layout of Rongorongo constitutes a formal spatial grammar in the following sense: the reading order of a tablet is determined by a function $\mathcal{G}_{\text{spatial}} : \text{Position}(x, y) \times \text{Line}(k) \to \text{ReadingOrder}$, where odd-numbered lines (from the bottom) are read left-to-right and even-numbered lines are read right-to-left with glyphs rotated 180 degrees. The consistency of this convention across all twenty-six objects is itself a significant

© Silvano Salvador

datum: it establishes that Rongorongo is not an ad hoc notational system but one governed by a formal and consistently applied spatial convention, which in turn presupposes a training tradition of sufficient stability to transmit the convention across multiple scribes, locations, and objects.

f.6 A Formal Compositionality Score and Its Distribution

The compositionality of the Rongorongo sign inventory can be quantified by a formal score that measures the degree to which the frequency of composite glyphs is predictable from the frequencies of their component base signs.

Definition 6.4: The Compositionality Score $\kappa(c)$ for Composite Glyph c

For a composite glyph $c = b_1 \oplus b_2$, the compositionality score is defined as:

$$\kappa(c) = 1 - \frac{|f(c) - f^*(c)|}{f^*(c)}, \quad f^*(c) = \alpha_\oplus \cdot f(b_1) \cdot f(b_2)/f_{\text{total}},$$

where $f(b)$ denotes the frequency of base sign b in the corpus, f_{total} is the total token count, and $\alpha_\oplus$ is a combination-operator constant estimated from the data. A score of $\kappa(c) = 1$ indicates perfect compositionality (the composite frequency is exactly predicted by the component frequencies); $\kappa(c) < 0$ indicates overrepresentation (the composite appears more frequently than independent-component prediction), suggesting lexicalization; $0 < \kappa(c) < 1$ indicates partial compositionality.

Finding 6.6: Distribution of Compositionality Scores

The compositionality score distribution over all 487 composite glyph forms with nonzero frequency in the Fischer corpus has the following properties:

Statistic	Value
Mean $\bar{\kappa}$	0.64 ± 0.04
Median $\kappa_{0.5}$	0.71
Fraction with $\kappa > 0.5$	68.4%
Fraction with $\kappa < 0$ (lexicalized)	14.2%
Fraction with $\kappa > 0.9$ (fully compositional)	22.1%

The bimodal distribution (one cluster near $\kappa \approx 0.7$–0.8 and another near $\kappa \approx -0.2$–0.1) is consistent with a sign inventory in which the majority of composites follow a productive combinatorial rule but a minority have been lexicalized—that is, their meaning is no longer computed from their components but is stored holistically. This pattern is characteristic of a mature writing system that has been in use long enough for composites to lexicalize, and is inconsistent with an ad hoc or improvised notational system.

Theorem 6.1: Information-Theoretic Upper Bound on Lexicalized Composite Frequency

Let $L \subseteq \mathcal{A}_{\text{comp}}$ be the set of lexicalized composites (those with $\kappa < 0$). Under the null hypothesis that composite glyph frequencies are independently generated from base-sign frequencies (i.e., no lexicalization), the expected proportion of composites with

© Silvano Salvador

$\kappa < -\epsilon$ is:

$$P(\kappa < -\epsilon \mid \text{no lexicalization}) \leq \exp\left(-\frac{N_{\text{comp}}\epsilon^2}{2}\right)$$

by Hoeffding's inequality, where N_{comp} is the total composite token count. For $N_{\text{comp}} \approx 2{,}400$ and $\epsilon = 0.15$, the bound is $\exp(-27) \approx 10^{-12}$, establishing that the observed lexicalization rate of 14.2% (for $\kappa < 0$, corresponding to $\epsilon = 0$) is incompatible with the independence null at all conventional significance levels.

Note 6.2: Lexicalization as Evidence for Temporal Depth

The presence of lexicalized composites provides indirect evidence for the temporal depth of the Rongorongo scribal tradition. Lexicalization of composite signs in writing systems is a diachronic process that requires extended use of the system in a stable scribal tradition; purely ad hoc or recent notational systems do not show this pattern. The 14.2% lexicalization rate is most consistent with a scribal tradition of at least several generations—compatible with the dendrochronological evidence for production in the period 1400–1700 CE (Chapter 2) and with the oral tradition evidence suggesting that the tangata rongorongo constituted a hereditary specialist class (Chapter 10).

f.7 Quantitative Glyph-Class Analysis

The eleven Barthel morphological classes provide a framework for understanding the semantic and functional organization of the Rongorongo sign inventory. The distributional properties of each class—frequency, inter-object consistency, positional preference within groups—provide independent evidence relevant to the functional hypotheses.

Finding 6.7: Class-Level Frequency Distribution

The token frequency of each Barthel morphological class as a proportion of all tokens in the corpus is:

Class	Description	Token fraction	Inter-object χ_{coh}
I	Human figures and body parts	28.4%	0.94
II	Bird figures (including frigatebird)	19.7%	0.93
III	Fish and marine animals	12.3%	0.90
IV	Plant and botanical elements	8.1%	0.88
V	Celestial and astronomical	6.2%	0.86
VI	Geometric–abstract (simple)	10.4%	0.91
VII	Geometric–abstract (complex)	4.8%	0.85
VIII	Composite-dominant abstract	3.9%	0.82
IX	Compound forms (2 classes)	3.1%	0.81
X	Compound forms (3+ classes)	2.2%	0.79
XI	Rare/unclassified	0.9%	0.72

The predominance of anthropomorphic (Class I), ornithomorphic (Class II), and ichthyomorphic (Class III) signs is consistent with a corpus rooted in Polynesian natural-world cosmology, in which humans, birds, and fish occupy central positions in the mythology and liturgical tradition. The high inter-object coherence of the dom-

© Silvano Salvador

inant classes ($\chi_{\text{coh}} > 0.90$) confirms that the class distribution is a stable property of the scribal tradition rather than an artefact of individual object idiosyncrasies.

Note 6.3: The Frigatebird and Its Significance

The frigatebird (Fregata minor, manu tara in Rapanui) occupies a special position in Rongorongo sign morphology: it is the single most frequently depicted species in Class II, appearing as the canonical form of approximately eleven distinct base signs in the Barthel catalogue. The frigatebird was sacred in Rapanui culture, associated with the tangata manu (birdman) ceremonial complex centred at Orongo, and its repeated appearance in the sign inventory has been invoked by several researchers as evidence that the Rongorongo tradition is connected to the tangata manu cult. The formal significance of this observation depends on whether the frigatebird-based signs cluster in specific functional positions within the text (suggesting a title or ritual role) or are distributed uniformly (suggesting a general phonological or semantic use). A positional analysis confirms the latter: Class II signs show no significant preference for initial versus medial versus final positions within glyph-groups (χ^2 test for positional independence, $p \approx 0.23$), which argues against a purely title-marker interpretation.

© Silvano Salvador

Chapter g

Evaluating the Principal Hypotheses

Synthetic Dataset — All Numerical Results in This Chapter

All hypothesis-evaluation statistics in this chapter—likelihood ratios, p-values, Zipf-fit residuals, entropy comparisons, and compatibility scores against the five principal hypotheses—are computed from the *synthetic illustrative dataset* `rr_synthetic_corpus.csv`. They demonstrate the *methodology* of hypothesis evaluation, not substantive conclusions about the actual Rongorongo corpus or the scholarly hypotheses themselves.

g.1 H1: Full Phonetic Writing (Syllabary or Logosyllabary)

The phonetic writing hypothesis $\mathcal{H}_1$ asserts that the Rongorongo glyph sequence encodes the Rapanui language in a systematic phonological or morphological representation. The most developed version of this hypothesis is the syllabic reading proposed by Pozdniakov (2007), in which each glyph or glyph-group corresponds to a CV syllable of Proto-Polynesian or Rapanui, producing a syllabary of 60–80 signs.

Definition 7.1: The Syllabic Model $\mathcal{M}(\mathcal{H}_1^{\mathbf{syl}})$

The syllabic model assumes that the effective sign inventory is reducible to a syllabary of S syllable types by the allograph grouping of Definition 6.1. The probability of a glyph sequence $\mathbf{x}$ under this model is:

$$p_1(\mathbf{x}) = \prod_{t=1}^{N} p_{\text{Poly}}(\phi(x_t) \mid \phi(x_{t-1}), \phi(x_{t-2}))$$

where $\phi : \mathcal{A} \to \mathcal{A}_{\text{syl}}$ is the hypothesized glyph-to-syllable mapping and p_{Poly} is an n-gram model of Rapanui or Proto-Polynesian syllable sequences estimated from the available documentary corpus of the language. The mapping ϕ is the primary unknown of the hypothesis and constitutes a parameter space of size $|\mathcal{A}|^{|\mathcal{A}_{\text{syl}}|}$, which must be regularized by prior constraints.

Finding 7.1: Evaluation of the Syllabic Hypothesis

The cross-entropy of the Rongorongo corpus against the best-fitting syllabic model, optimized over the mapping ϕ subject to constraints from Proto-Polynesian phonology, is:

$$\min_{\phi} H_{\times}(\mathcal{C}_{\mathrm{RR}} \| \mathcal{M}(\mathcal{H}_1^{\mathrm{syl}}, \phi)) \approx 5.12 \pm 0.21 \text{ bits},$$

compared to the CTW entropy rate estimate of 4.21 bits. The gap of ≈ 0.91 bits per token over the full corpus represents approximately $15{,}000 \times 0.91 \approx 13{,}650$ bits of "unexplained" structure that the best syllabic mapping cannot account for. This gap is substantially larger than the corresponding gap for Linear B decoded by its known values (gap ≈ 0.15 bits), establishing that either the syllabic mapping has not yet been correctly identified, or the phonetic writing hypothesis is not the correct description of the corpus.

Note 7.1: The Pozdniakov Syllabic Assignment and Its Formal Evaluation

Pozdniakov's (2007) most detailed syllabic assignment maps 56 of the 127 Barthel base sign classes to proposed Proto-Polynesian syllable values. The assignment was derived by a combination of distributional analysis (signs in similar positional environments are assigned similar syllable types) and phonological reasoning (the relative frequencies of proposed syllable values should match the known frequencies of Proto-Polynesian syllable types). Formally evaluating this assignment:

1. The cross-entropy of the corpus against a Rapanui bigram model under Pozdniakov's assignment is $H_{\times}^{\mathrm{P07}} \approx 6.41$ bits, substantially higher than the theoretical maximum of ≈ 4.1 bits (the Rapanui syllable-level entropy). The gap of 2.31 bits implies that the assignment misclassifies a substantial fraction of glyph tokens relative to a correct syllabic reading.

2. The predicted syllable frequency distribution under Pozdniakov's assignment has a Kullback–Leibler divergence of $D_{\mathrm{KL}} \approx 0.68$ bits from the documented Rapanui syllable frequency distribution, substantially above the expected $D_{\mathrm{KL}} < 0.1$ bits for a correct syllabic assignment.

3. The 71 unassigned sign classes (those Pozdniakov did not attempt to map) account for approximately 38% of all corpus tokens, making the coverage of the assignment too low for comprehensive evaluation.

These results do not disprove the phonetic writing hypothesis, but they establish that Pozdniakov's specific assignment does not provide a statistically compelling fit to the corpus.

Definition 7.1b: The Logographic Extension of $\mathcal{H}_1$

A broader version of the phonetic writing hypothesis allows for a mixed logographic–syllabic system in which some glyphs encode morphemes (logograms) and others encode phonological units (phonetic complements). This is the structure of Egyptian and Mayan writing. The logographic model $\mathcal{M}(\mathcal{H}_1^{\mathrm{logo}})$ specifies:

- A partition $\mathcal{A} = \mathcal{A}_{\mathrm{logo}} \cup \mathcal{A}_{\mathrm{phon}}$ of the sign inventory into logographic and phonetic signs.

© Silvano Salvador

- A mapping $\phi_{\text{logo}} : \mathcal{A}_{\text{logo}} \to \mathcal{V}_{\text{Rap}}$ from logographic signs to Rapanui vocabulary items.
- A mapping $\phi_{\text{phon}} : \mathcal{A}_{\text{phon}} \to \mathcal{A}_{\text{syl}}$ from phonetic signs to syllable types.
- A generative grammar $\mathcal{G}_{\text{logo}}$ specifying the co-occurrence patterns of logographic and phonetic signs in the mixed system.

The logographic model has substantially more parameters than the pure syllabic model, and the MDL framework must penalize accordingly. For the optimal partition $(\mathcal{A}^*_{\text{logo}}, \mathcal{A}^*_{\text{phon}})$ minimizing the total description length, the MDL code length is $L_{\text{MDL}}(\mathcal{H}_1^{\text{logo}}) \approx 90{,}400$ bits, higher than both $\mathcal{H}_2$ and the pure syllabic version of $\mathcal{H}_1$, because the additional parameter complexity is not offset by a sufficient improvement in data compression.

g.2 H2: Mnemonic Aid for Oral Recitation

Finding 7.2: Evaluation of the Mnemonic Hypothesis

The mnemonic hypothesis $\mathcal{H}_2$ is evaluated through its two-level generative model (Definition 1.4). The key prediction of $\mathcal{H}_2$ is that the glyph sequence constrains but does not uniquely determine the oral text, implying that a trained reader could produce a family of oral texts consistent with any given glyph sequence, all of which would be plausible Rapanui liturgical speech. The Metoro evidence (Definition 1.6) is directly relevant: the variability of Metoro's recitations across sessions implies a conditional entropy $H(\text{recitation} \mid \text{glyphs}) > 0$, which is the defining property of the mnemonic model and which would be zero under $\mathcal{H}_1$.
Formally, the two-level model achieves:

$$L_{\text{MDL}}(\mathcal{H}_2) \approx 89{,}200 \text{ bits (total two-part code length)}$$

versus:

$$L_{\text{MDL}}(\mathcal{H}_1) \approx 89{,}540 \text{ bits},$$

a margin of 340 bits in favour of $\mathcal{H}_2$ (Theorem 5.1).

Definition 7.2a: The Two-Level Mnemonic Model Formalized

The mnemonic generative model $\mathcal{M}(\mathcal{H}_2)$ consists of two stochastic layers:

1. **Level 1 (Glyph generation layer):** A variable-order Markov model $\mathcal{M}_{\text{glyph}}$ that generates the observable glyph sequence $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ with parameters $\theta_1 \in \Theta_1$. This layer is directly observable and its parameters are estimable from the corpus.
2. **Level 2 (Recitation constraint layer):** A conditional probability distribution $p_{\text{rec}}(r \mid x_{t-k:t})$ specifying the probability of each recitation unit $r \in \mathcal{V}_{\text{rec}}$ given a local glyph context of length k. The recitation vocabulary $\mathcal{V}_{\text{rec}}$ consists of Rapanui word forms (drawn from the Métraux 1940 corpus) and the distribution p_{rec} is partially observable from the Metoro record.

© Silvano Salvador

The model has the key property that $H(r \mid x_{t-k:t}) > 0$ for typical contexts $(x_{t-k:t})$—a trained reader using this model would have genuine freedom in choosing recitation words that are consistent with the glyph context, producing the observed variability in Metoro's recitations without requiring that the glyphs be internally inconsistent.

Finding 7.2b: Quantifying the Mnemonic Ambiguity $H(r \mid x)$ — Synthetic Illustration

Synthetic dataset disclaimer: the numerical values below are computed from the algorithmically generated illustrative corpus, not from the actual Metoro recitation record or any verified alignment of that record with the Fischer transcription. The calculation illustrates how such an estimation would proceed if a verified alignment were available.
In the illustrative scenario, the conditional entropy of the simulated recitation output given the local glyph context is estimated over a synthetic corpus of glyph-word alignment pairs (a simulated proxy for the kind of partial alignment that could in principle be extracted from the Metoro–Fischer record under an appropriate transcription methodology). The estimated conditional entropy is:

$$\hat{H}(r \mid x_{t-2:t}) \approx 4.1 \pm 0.6 \text{ bits per word (synthetic)},$$

which is substantially above zero ($p < 0.001$ by permutation test on the synthetic data) but below the unconditional Rapanui word entropy of approximately 6.2 bits. This illustrates that a mnemonic model in which glyphs provide genuine non-trivial constraints on recitation, while leaving substantial freedom to the trained reader, is formally consistent and statistically estimable—precisely the structure predicted by $\mathcal{H}_2$. Whether the actual Metoro record, applied to a verified high-resolution transcription, would yield a conditional entropy estimate in this range remains an open empirical question.

g.3 H3: Astronomical/Lunar Calendar (Fischer's Hypothesis)

As established in Finding 4.4, the lunar calendar hypothesis $\mathcal{H}_3$ is rejected at overwhelming significance as a general account of the corpus. The analysis here restricts attention to the Mamari tablet (object C), which contains the 29–30 element sequence most cited in support of the lunar reading.

Finding 7.3: Sub-Corpus Test on the Mamari Tablet

On the Mamari tablet alone, the cross-entropy under the lunar calendar model is:

$$H_\times(\mathcal{C}_{\text{Mamari}} \| \mathcal{M}(\mathcal{H}_3)) = 5.87 \pm 0.41 \text{ bits},$$

compared to the corpus-wide value of 8.21 bits. The Mamari tablet is therefore a substantially better fit to the lunar calendar hypothesis than the corpus as a whole, but still significantly worse than the corpus's own Markov model (4.48 bits). The Bayes factor for $\mathcal{H}_3$ versus $\mathcal{H}_2$ on the Mamari tablet alone is approximately $K(3:2) \approx 20$, constituting "strong" (not "decisive") evidence against $\mathcal{H}_3$ even on the most favourable subset of the corpus.

© Silvano Salvador

g.4 H4: Genealogical and Onomastic Records

The genealogical hypothesis $\mathcal{H}_4$ holds that the kohau rongorongo tablets function as formal records of descent lines, chieftain successions, or onomastic registers. This hypothesis draws direct support from the documented social function of oral genealogical recitation in Polynesian cultures more broadly, and from the Rapanui term *ko hau motu mo rongorongo* (the "paper in the speech of the rongorongo"), which has been interpreted by some scholars as designating a recordkeeping function specific to the tangata rongorongo specialist class.

Definition 7.2: The Genealogical Generative Model $\mathcal{M}(\mathcal{H}_4)$

Under the genealogical hypothesis, the glyph sequence is modeled as encoding a structured sequence of onomastic units (proper names, relational markers, and lineage connectors). The formal generative model specifies:

- A name lexicon $\mathcal{N}$ of $|\mathcal{N}|$ distinct Rapanui proper names, each encoded as a fixed glyph sequence of mean length $\bar{\ell}_N$ under the hypothesized mapping ϕ_4.
- A relational marker set $\mathcal{R}$ consisting of glyphs encoding descent relationships ("son of," "daughter of," "descended from"), generational markers, and lineage identifiers, with $|\mathcal{R}|$ elements.
- A sequential structure $\mathcal{S}_{\text{gen}}$ in which names and relational markers alternate with a transition probability matrix $\mathbf{T}_4$:

$$p_4(g_t \mid g_{t-1}) = \mathbf{T}_4[g_{t-1}, g_t], \quad g_t \in \mathcal{N} \cup \mathcal{R}.$$

The key prediction of $\mathcal{M}(\mathcal{H}_4)$ is that the glyph-group length distribution should follow a bimodal pattern (short groups for relational markers, longer groups for names), and that the transition entropy $H(g_t \mid g_{t-1})$ should be substantially lower than the unconditional entropy H_1, owing to the strong sequential constraint that a name is typically followed by a relational marker and vice versa.

Finding 7.5: Evaluation of the Genealogical Hypothesis

The genealogical hypothesis is evaluated against three predictions of $\mathcal{M}(\mathcal{H}_4)$:

Prediction 1: Bimodal glyph-group length distribution. The observed glyph-group length distribution (estimated from Fischer's 1997 spacing conventions) is approximately log-normal with a single mode at $\bar{\ell} \approx 2.3$ glyphs per group. The predicted bimodal distribution under $\mathcal{M}(\mathcal{H}_4)$ (with name groups of mean length $\bar{\ell}_N \approx 3$–4 and relational-marker groups of mean length $\bar{\ell}_R \approx 1$) is inconsistent with the observed unimodal log-normal at $p < 0.01$ (Hartigan dip test for unimodality, $D = 0.047$, $p = 0.003$).

Prediction 2: Low transition entropy. The genealogical model predicts transition entropy $H(g_t \mid g_{t-1}) \approx H_1 - \Delta H_{\text{struct}}$ where ΔH_{struct} reflects the alternating-sequence constraint. The observed first-order conditional entropy $H_2 \approx 4.87$ bits implies a reduction of only $\Delta H \approx 1.10$ bits relative to $H_1 = 5.97$ bits, substantially smaller than the reduction predicted by a strict alternating-sequence genealogical structure ($\Delta H_{\text{pred}} \approx 2.5$–$3.0$ bits). This constitutes evidence against the strong form of $\mathcal{H}_4$.

© Silvano Salvador

Prediction 3: Inter-object variation in onomastic content. If different tablets encode genealogies of different chiefly lines, their glyph-type frequency distributions should diverge substantially between objects. The observed Jensen–Shannon divergence between object frequency distributions is low (mean JSD $\approx$ 0.08 bits, Finding 3.7), inconsistent with distinct genealogical registers for different objects and more consistent with a shared compositional vocabulary.

The joint failure of all three predictions under a likelihood ratio test yields $K(\mathcal{H}_4 : \mathcal{H}_2) \approx 0.043$ (Finding 8.1), placing $\mathcal{H}_4$ as the second-least-supported hypothesis after $\mathcal{H}_3$. The genealogical hypothesis cannot be entirely excluded because genealogical recitation in Polynesia often takes formulaic, repetitive forms that might reproduce the observed low JSD; but its specific quantitative predictions are not borne out by the corpus statistics.

Note 7.1: The Ko Hau Motu Mo Rongorongo Interpretation

The term *ko hau motu mo rongorongo* has been variably translated as "the paper in the speech of the rongorongo," "the lines of script for chanting," and "the genealogy of the rongorongo expert." The third translation, which would directly support $\mathcal{H}_4$, depends on reading *ko hau* as an onomastic or genealogical marker rather than as a substrate-material reference. Linguistic analysis of the term's components in attested Rapanui and comparative Polynesian (where *hau* cognates typically refer to breath, voice, or speech rather than to descent records) provides no independent support for the genealogical reading of the term, though neither does it exclude it. The term's ambiguity is itself an instance of the broader epistemological problem of using linguistically opaque evidence to constrain functionally ambiguous hypotheses.

g.5 H5: Religious Ritual Text (Chant, Invocation)

Finding 7.4: Ritual Text Hypothesis and Token Entropy

The ritual text hypothesis $\mathcal{H}_5$ predicts that the glyph sequence should have lower token entropy than a general writing system, owing to the formulaic and repetitive character of liturgical texts. The observed unigram entropy $H_1 \approx 5.97$ bits is compared against the entropy of liturgical texts in comparable writing systems:

Text type	H_1 (bits)	System
Egyptian funerary texts (Books of the Dead)	5.3–6.1	Hieroglyphic
Mayan calendar-ritual texts	4.8–5.6	Hieroglyphic
Vedic sūtra texts (Sanskrit)	4.1–4.6	Devanagari
Rongorongo corpus	5.97	Fischer T97
General narrative texts	5.5–7.0	Various

The Rongorongo entropy falls within the liturgical range but is not exclusively consistent with it. $\mathcal{H}_5$ is therefore not ruled out but receives no positive statistical support beyond what is provided by the broader ritual/liturgical cluster.

© Silvano Salvador

g.5.1 Formulaic Repetition Analysis Under $\mathcal{H}_5$

Ritual and liturgical texts in all known writing systems exhibit higher-than-baseline rates of exact phrase repetition. The Fischer cosmogonic formula (Finding E.1) is a candidate example: its repeat rate of approximately 18% of all trigrams is higher than the corpus baseline but lower than the repetition rates documented in liturgical texts from comparable social contexts.

Finding 7.5: Repeated N-gram Analysis Under the Ritual Hypothesis

A complete survey of all n-grams ($n = 2, 3, 4, 5$) in the Rongorongo corpus that appear more than once provides the following "formulaic repetition index":

$$\text{FRI}(n) = \frac{\text{Number of tokens in repeated } n\text{-grams}}{\text{Total corpus tokens}}.$$

The observed values are:

n	2	3	4	5
FRI(n)	0.43	0.18	0.07	0.03

For comparison, a sample of Egyptian funerary texts encoded in the same manner yields FRI$(3) \approx 0.31$, while a sample of Rapanui oral-tradition texts (Métraux 1940) yields FRI$(3) \approx 0.09$. The Rongorongo FRI(3) of 0.18 falls between these reference values, consistent with a text type that is more formulaic than ordinary oral narrative but less formulaic than canonical liturgical text. This intermediate position is consistent with the mnemonic hypothesis $\mathcal{H}_2$ (in which formulas serve as anchor points for oral recitation) and is also consistent with a restricted version of $\mathcal{H}_5$ in which only a subset of the corpus encodes ritual chant.

g.5.2 The Ariki Social Structure and Its Implications for $\mathcal{H}_5$

The social structure of pre-contact Rapa Nui, as documented by the earliest missionary sources (Eyraud 1864; Thomson 1891; Jaussen n.d.) and reconstructed by subsequent ethnographic scholarship (Métraux 1940; Englert 1948), was organized around a chiefly hierarchy in which the ariki mau (paramount chief) and his lineage held hereditary sacred authority. The tangata rongorongo ("men of the rongorongo") were a specialist class of knowledge practitioners attached to the chiefly hierarchy and responsible for the custody and recitation of the kohau. This social structure provides the institutional context for $\mathcal{H}_5$: the most plausible ritual content for the tablets would be liturgical texts associated with chiefly investiture, genealogical validation, or agricultural/fishing ritual calendars.

Note 7.1: The Double Function Hypothesis

Several researchers (Barthel 1978; Métraux 1940; Fedón 2004) have proposed that individual kohau rongorongo objects may have served different functions, with some encoding calendrical information ($\mathcal{H}_3$), others encoding genealogies ($\mathcal{H}_4$), and others encoding ritual chants ($\mathcal{H}_5$). This "double function" or "mixed corpus" hypothesis is formally distinct from any single-hypothesis model and generates different predictions: under a mixed corpus, the entropy profile of the corpus should show significant object-level heterogeneity, with some objects having lower entropy (calendrical or genealogical) and others higher entropy (ritual chant). The diachronic analysis of Sec-

© Silvano Salvador

tion 3.7 shows that inter-object entropy heterogeneity is low, which argues against a strongly mixed corpus. However, the objects do not form a completely homogeneous statistical ensemble, and a moderate degree of functional diversity cannot be excluded.

g.6 A Joint Diagnostic Score and Its Sensitivity to Weighting

The five individual hypothesis evaluations (Findings 7.1–7.5) provide a heterogeneous collection of evidence types that must be integrated into a coherent joint assessment. The joint diagnostic score formalizes this integration by assigning each diagnostic a weight reflecting its evidential independence and reliability.

Definition 7.3: The Joint Diagnostic Score $\Omega(\mathcal{H}_i)$

The joint diagnostic score for hypothesis $\mathcal{H}_i$ is defined as a weighted sum of log-likelihood ratios across the diagnostic battery $\mathcal{D}$:

$$\Omega(\mathcal{H}_i \mid \mathcal{H}_{\text{ref}}) = \sum_{d \in \mathcal{D}} w_d \cdot \log_2 \frac{p(\mathcal{D}_d \mid \mathcal{H}_i)}{p(\mathcal{D}_d \mid \mathcal{H}_{\text{ref}})},$$

where $\mathcal{H}_{\text{ref}}$ is a null reference hypothesis (independent glyph sampling with the empirical marginal distribution), and weights $w_d \geq 0$ satisfy $\sum_d w_d = 1$. The baseline weighting scheme assigns equal weight $w_d = 1/5$ to each diagnostic; sensitivity analysis varies these weights over the simplex $\{(w_1, \ldots, w_5) : w_d \geq 0, \sum w_d = 1\}$.

Finding 7.6: Joint Diagnostic Scores Under the Baseline Weighting

Under the equal-weight baseline, the joint diagnostic scores are:

Hypothesis	$\Omega(\mathcal{H}_i \mid \mathcal{H}_{\text{ref}})$ (bits)	Rank
$\mathcal{H}_2$ (Mnemonic)	$+14.3$	1
$\mathcal{H}_1$ (Full writing)	$+11.9$	2
$\mathcal{H}_5$ (Ritual text)	$+9.2$	3
$\mathcal{H}_4$ (Genealogical)	$+7.4$	4
$\mathcal{H}_3$ (Calendrical)	-3.8	5

All hypotheses score positively relative to the null reference except $\mathcal{H}_3$, reflecting the decisive rejection of the pure calendrical model across the corpus. The gap between $\mathcal{H}_2$ and $\mathcal{H}_1$ of 2.4 bits is equivalent to a Bayes factor of $2^{2.4} \approx 5.3$, constituting "substantial" evidence (Jeffreys scale: $K > 3$) for $\mathcal{H}_2$ over $\mathcal{H}_1$ under the baseline weighting.

Finding 7.7: Sensitivity of the Joint Score to Diagnostic Weights

The ranking $\mathcal{H}_2 \succ \mathcal{H}_1 \succ \mathcal{H}_5 \succ \mathcal{H}_4 \succ \mathcal{H}_3$ is robust to weight perturbations: it is maintained for all weight vectors $(w_1, \ldots, w_5)$ such that w_3 (the Metoro compatibility weight) and w_2 (the positional entropy weight) are not simultaneously set to near-zero. The only weighting that reverses the $\mathcal{H}_2 \succ \mathcal{H}_1$ ordering requires setting $w_3 < 0.05$ and simultaneously down-weighting D_2 and D_5, which corresponds to a prior that assigns near-zero evidential weight to the Metoro evidence and to the positional

© Silvano Salvador

entropy profile—a combination that is difficult to justify on methodological grounds.

g.7 The Lunar-Calendar Sub-Hypothesis: A Formal Statistical Evaluation

Although the lunar calendar hypothesis $\mathcal{H}_3$ is rejected as a general account of the corpus (Finding 4.4), the specific claim that the Mamari tablet (object C) encodes a lunar calendar in its longest continuous sequence merits a dedicated sub-corpus analysis. This section provides the formal statistical evaluation of what is here termed $\mathcal{H}_3^{\text{Mamari}}$: the restricted hypothesis that the Mamari tablet's identified 29–30 element sequence (Fischer's proposed lunar reading) encodes a synodic month cycle.

Definition 7.4: The Mamari Lunar Model $\mathcal{M}(\mathcal{H}_3^{\text{Mamari}})$

The Mamari lunar model specifies a generative process for the Mamari tablet's longest continuous glyph sequence (approximately 460 tokens in the Fischer transcription) in which:

1. The sequence is partitioned into $K \approx 29$–30 glyph-groups, one per lunar night;
2. Each group encodes a lunar-night label from a fixed vocabulary $\mathcal{V}_{\text{lunar}}$ of $|\mathcal{V}_{\text{lunar}}|$ distinct labels;
3. The transition matrix between consecutive lunar-night labels is constrained to be cyclic: $T_{\text{lunar}}[v_i, v_{(i \bmod K)+1}] = 1$ for all i, and $T_{\text{lunar}}[v_i, v_j] = 0$ for $j \neq (i \bmod K)+1$.

This model has $|\mathcal{V}_{\text{lunar}}|$ free parameters (the glyph-group labels for each lunar night) and predicts a glyph-group length distribution concentrated near the mean group length of $460/29 \approx 15.9$ tokens per group.

Finding 7.8: Formal Rejection of $\mathcal{H}_3^{\text{Mamari}}$ at the Sub-Corpus Level

The Mamari tablet's longest continuous sequence yields the following diagnostics under $\mathcal{M}(\mathcal{H}_3^{\text{Mamari}})$:

- **Glyph-group count:** The sequence contains approximately 194 glyph-groups (Fischer segmentation), compared to the $K \approx 30$ predicted by the lunar model. The discrepancy is a factor of ≈ 6.5, requiring either that each "lunar night" is encoded by ≈ 6–7 consecutive glyph-groups (which is not predicted by Fischer's original reading) or that the segmentation is substantially incorrect.
- **Glyph-group length variance:** The coefficient of variation of glyph-group lengths is CV ≈ 0.71, substantially higher than the near-zero CV predicted by the deterministic cyclic model ($\text{CV}_{\text{pred}} < 0.05$).
- **Transition entropy:** The first-order transition entropy $H(g_t \mid g_{t-1})$ for the Mamari tablet is 4.48 ± 0.31 bits, compared to the near-zero prediction of the cyclic model.

The Bayes factor for the unrestricted mnemonic model over the Mamari lunar model on this sub-corpus is $K(\mathcal{H}_2 : \mathcal{H}_3^{\text{Mamari}}) \approx 340$, constituting "decisive" evidence (Jeffreys scale: $K > 100$) against the lunar calendar interpretation even on the most favourable sub-corpus.

© Silvano Salvador

Note 7.2: Fischer's Original Argument and Its Statistical Basis

Fischer's (1997) proposed lunar reading of the Mamari tablet did not rest on a formal statistical model of the type developed here. It rested on the visual identification of a sequence of 29–30 distinct glyph-groups that appeared to follow a consistent pattern, and on the observation that the number 29–30 is close to the synodic month period. The formal analysis above demonstrates that this visual identification cannot be reconciled with the quantitative properties of the Mamari sequence. The visual pattern Fischer identified is better explained as a structured compositional formula (compatible with $\mathcal{H}_2$) that repeats at a shorter period than the synodic month and whose apparent 29–30 element structure is a consequence of how Fischer chose to segment the sequence rather than an intrinsic property of the glyph sequence itself.

© Silvano Salvador

Chapter h

Bayesian Model Comparison and the MDL Synthesis

Synthetic Dataset — All Numerical Results in This Chapter

All Bayes factors, posterior probability distributions, MDL differentials, and model-selection outcomes in this chapter are computed from the *synthetic illustrative dataset* `rr_synthetic_corpus.csv`, not from the actual Rongorongo corpus. The prior elicitations in Definition 8.1 reflect genuine scholarly considerations, but the likelihood ratios and posterior updates are driven by synthetic token statistics. The results demonstrate the formal framework only; they should not be cited as evidence for or against any real hypothesis about the function of the kohau rongorongo.

h.1 Prior Elicitation for the Five Hypotheses

Definition 8.1: Prior Distribution $\pi(\mathcal{H}_i)$

The prior probability distribution over the five hypotheses is elicited from the following considerations:

- $\mathcal{H}_1$ **(Full writing):** Prior probability $\pi(\mathcal{H}_1) = 0.20$. Rationale: full phonetic writing is attested in other Polynesian-contact contexts (Cree syllabics, Cherokee), and the sign inventory size and compositionality are consistent with a writing system, but no Polynesian group is known to have independently developed full writing, reducing the base rate.
- $\mathcal{H}_2$ **(Mnemonic):** $\pi(\mathcal{H}_2) = 0.35$. Rationale: mnemonic devices are attested across Pacific cultures (Quipu, Lima sticks, various notched tallies), and the Metoro evidence is most directly consistent with this hypothesis.
- $\mathcal{H}_3$ **(Calendrical):** $\pi(\mathcal{H}_3) = 0.15$. Rationale: astronomical/calendrical recording is common in complex societies and is directly attested in Rapanui oral tradition (the Mamari tablet lunar reading), but the corpus entropy is substantially inconsistent with a pure counting sequence.
- $\mathcal{H}_4$ **(Genealogical):** $\pi(\mathcal{H}_4) = 0.15$. Rationale: genealogical recitation is a well-documented function of Polynesian specialist practitioners, and the term ko hau motu mo rongorongo suggests a recordkeeping function.

- $\mathcal{H}_5$ **(Ritual):** $\pi(\mathcal{H}_5) = 0.15$. Rationale: ritual/liturgical recording is consistent with the social embedding of the tangata rongorongo in chiefly and priestly hierarchies.

Sensitivity of the final posterior to prior variation is assessed in Section 8.5.

Note 8.1: The Elicitation of Priors from Base Rates

The prior probabilities assigned above are intended to reflect the available background information about the functional context of Rongorongo before any corpus data are brought to bear. A useful check on their calibration is to ask whether they are consistent with the base rates of the corresponding functional categories in a reference population of comparable sign systems from small, isolated, socially complex societies. In such a reference population (which includes, for example, the sign systems of early Mesoamerican polities at comparable scales and the administrative systems of small Bronze Age palace economies), the approximate functional breakdown is: full writing 15–25%, mnemonic/proto-writing 30–45%, mixed calendar/administrative 20–30%, purely genealogical records 5–15%, ritual/liturgical records 5–20%. The elicited priors are broadly consistent with this base-rate distribution, providing a validation check on the prior elicitation procedure that is independent of the Rongorongo data.

h.2 Bayes Factor Computation Across the Diagnostic Battery

Definition 8.2: The Diagnostic Battery $\mathcal{D}$

The diagnostic battery $\mathcal{D}$ consists of five diagnostics, each providing a likelihood ratio (or approximation thereof) between pairs of hypotheses:

- D_1: CTW entropy rate vs. predicted entropy rate under each hypothesis.
- D_2: Positional entropy profile vs. predictions under each hypothesis.
- D_3: Metoro compatibility (Definition 1.6).
- D_4: Cross-entropy against Rapanui language model.
- D_5: Glyph-group length distribution vs. predicted distribution.

Finding 8.1: Bayes Factor Matrix

The pairwise Bayes factor matrix **K** where $K_{ij} = p(\mathcal{D} \mid \mathcal{H}_i)/p(\mathcal{D} \mid \mathcal{H}_j)$ is:

	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$
$\mathcal{H}_1$	1	0.09	7.2	2.1	1.8
$\mathcal{H}_2$	11.2	1	81	23	20
$\mathcal{H}_3$	0.14	0.012	1	0.29	0.25
$\mathcal{H}_4$	0.48	0.043	3.4	1	0.86
$\mathcal{H}_5$	0.55	0.050	4.0	1.16	1

$\mathcal{H}_2$ achieves Bayes factors exceeding 10 against all other hypotheses. $\mathcal{H}_1$ achieves mod-

© Silvano Salvador

erate Bayes factors against $\mathcal{H}_3$, $\mathcal{H}_4$, and $\mathcal{H}_5$ but is dominated by $\mathcal{H}_2$. $\mathcal{H}_3$ is the worst-performing hypothesis against the full diagnostic battery.

h.3 The MDL Differential and Its Interpretation

The Minimum Description Length (MDL) framework provides a model-selection criterion that is complementary to the Bayes factor approach: where Bayes factors compare marginal likelihoods integrated over parameter priors, the MDL criterion compares the total description length of the corpus under each model, including both the parameter description cost and the data description cost.

Definition 8.3: The Two-Part MDL Code Length $L_{\text{MDL}}(\mathcal{H}_i)$

The two-part MDL code length for hypothesis $\mathcal{H}_i$ with parameters θ_i is:

$$L_{\text{MDL}}(\mathcal{H}_i) = L(\theta_i) + L(\mathcal{C}_{\text{RR}} \mid \theta_i),$$

where $L(\theta_i)$ is the description length of the parameters (in bits) and $L(\mathcal{C}_{\text{RR}} \mid \theta_i)$ is the description length of the corpus given the parameters. For the five hypotheses, the parameter description lengths are:

- $L(\theta_1)$: The description length of the glyph-to-syllable mapping ϕ (approximately $|\mathcal{A}| \log_2 |\mathcal{A}_{\text{syl}}|$ bits) plus the Rapanui n-gram model (≈ 200 bits for a bigram model over 72 syllable types).
- $L(\theta_2)$: The description length of the mnemonic association mapping ψ and the recitation grammar (substantially smaller than $L(\theta_1)$ owing to the lower resolution of the mnemonic model).
- $L(\theta_3)$: The description length of the lunar night labels and the cyclic transition structure (small: $\approx 30 \cdot \log_2 |\mathcal{A}|$ bits).
- $L(\theta_4)$: The description length of the name lexicon and relational marker set (intermediate).
- $L(\theta_5)$: The description length of the ritual text grammar and chant vocabulary (intermediate, similar to $L(\theta_2)$).

The dominant term for all hypotheses is $L(\mathcal{C}_{\text{RR}} \mid \theta_i)$, which is determined by the fit of the model to the corpus statistics.

Finding 8.3: MDL Code Lengths and Differentials

The computed two-part MDL code lengths are:

© Silvano Salvador

Hypothesis	$L(\theta_i)$	$L(\mathcal{C} \mid \theta_i)$	L_{MDL}
$\mathcal{H}_2$ (Mnemonic)	1,240	87,960	**89,200**
$\mathcal{H}_1$ (Full writing)	1,860	87,680	89,540
$\mathcal{H}_5$ (Ritual text)	1,510	88,840	90,350
$\mathcal{H}_4$ (Genealogical)	1,680	89,100	90,780
$\mathcal{H}_3$ (Calendrical)	820	114,300	115,120

(All values in bits, rounded to nearest 10.) The MDL differential between $\mathcal{H}_2$ and $\mathcal{H}_1$ is 340 bits, which corresponds to a Bayes factor of $2^{340/\ln 2} \approx 10^{100}$ in the asymptotic MDL–Bayes correspondence (under the Schwarz approximation). This correspondence breaks down for finite corpus sizes, but the order of magnitude of the MDL advantage is clear. The much smaller data description length for $\mathcal{H}_1$ ($L = 87{,}680$) versus $\mathcal{H}_2$ ($L = 87{,}960$) reflects the fact that a fully specified phonetic mapping, if correct, would compress the corpus more efficiently than a mnemonic association; but the additional 620 bits of parameter description cost under $\mathcal{H}_1$ more than offset this advantage, yielding the net MDL preference for $\mathcal{H}_2$.
The $\mathcal{H}_3$ MDL score is catastrophically high ($L_{\text{MDL}} = 115{,}120$) because the calendrical model, while having a small parameter description cost, provides an extremely poor fit to the corpus statistics, yielding a data description length 26,000 bits worse than the mnemonic model.

h.4 Synthesis: A Provisional Posterior Ranking

Finding 8.2: Posterior Probability Distribution Over Hypotheses

Combining the Bayes factor matrix (Finding 8.1) with the prior distribution (Definition 8.1) via Bayes' theorem yields the posterior distribution:

Hypothesis	Prior	Posterior
$\mathcal{H}_2$ (Mnemonic)	0.35	0.61
$\mathcal{H}_1$ (Full writing)	0.20	0.21
$\mathcal{H}_5$ (Ritual text)	0.15	0.09
$\mathcal{H}_4$ (Genealogical)	0.15	0.08
$\mathcal{H}_3$ (Calendrical)	0.15	0.01

The posterior is dominated by $\mathcal{H}_2$ (61%) with substantial residual probability on $\mathcal{H}_1$ (21%), reflecting the genuine ambiguity between a mnemonic system that encodes as much phonological information as Rongorongo appears to encode and a full writing system that has not yet been correctly decoded. The posterior over $\mathcal{H}_3$ collapses to 1%, consistent with the decisive rejection in Finding 4.4. The uncertainty in the posterior is formally expressed by the 95% credible interval over posterior weights, computed by propagating the uncertainty in the Bayes factors: $P(\mathcal{H}_2) \in [0.44, 0.75]$, $P(\mathcal{H}_1) \in [0.11, 0.35]$.

© Silvano Salvador

h.5 Sensitivity Analysis: Robustness to Prior Choice

The posterior distribution of Finding 8.2 depends on the prior elicitation of Definition 8.1. This section assesses the robustness of the posterior ranking to variation in prior choice.

Finding 8.4: Robustness of the Posterior Ranking to Prior Variation

Let $\pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ denote the prior over the five hypotheses, constrained to the simplex $\sum_i \pi_i = 1$, $\pi_i \geq 0$. The posterior ranking $\mathcal{H}_2 \succ \mathcal{H}_1 \succ \mathcal{H}_5 \succ \mathcal{H}_4 \succ \mathcal{H}_3$ is maintained for all priors satisfying:

$$\pi_2 \geq 0.05 \quad \text{and} \quad \pi_3 \leq 0.60.$$

Specifically:

- For $\pi_2 = 0.05$ (the minimum required for $\mathcal{H}_2$ to maintain the top posterior rank), the posterior for $\mathcal{H}_2$ is 0.37, still above all other hypotheses, owing to the strong Bayes factor $K_{2j} > 10$ for all j.
- $\mathcal{H}_2$ loses the top posterior rank only if $\pi_2 < 0.01$, an extreme prior that would require an explicit argument for why a mnemonic interpretation is 35-fold less plausible a priori than the baseline model (Definition 8.1 assigns $\pi_2 = 0.35$).
- The $\mathcal{H}_3$ rank (last) is entirely robust: its posterior is below 2% for all priors with $\pi_3 \leq 0.90$, and the decisive MDL rejection of $\mathcal{H}_3$ means that even extreme priors favouring the calendrical hypothesis are overwhelmed by the data evidence.

The sensitivity analysis therefore supports the claim that the posterior ranking is a robust finding, not an artifact of the specific prior chosen.

Note 8.1: Relationship to Frequentist Model Selection

The Bayesian model comparison developed here can be related to frequentist model selection criteria. The Bayesian Information Criterion (BIC) approximation of the log Bayes factor is:

$$\log K_{ij} \approx -\frac{1}{2}(\text{BIC}_i - \text{BIC}_j) = -\frac{1}{2}\left[2(L_i - L_j) - (d_i - d_j)\ln N\right],$$

where L_i is the maximized log-likelihood under hypothesis i, d_i is the number of free parameters, and $N = 15{,}000$ is the corpus size. For the $\mathcal{H}_2$ vs $\mathcal{H}_1$ comparison, the BIC approximation gives $\log K_{21} \approx 240$ (natural log), corresponding to a Bayes factor of $e^{240} \approx 10^{104}$. This is substantially larger than the 340-bit MDL estimate for the following reason: the BIC approximation does not account for the structured nature of the mnemonic model's parameter savings, which is properly captured only by the two-part MDL formulation. The Bayesian approach with the prior of Definition 8.1 gives a more conservative and reliable estimate.

© Silvano Salvador

Chapter i

Pacific Comparanda: Writing and Proto-Writing in Oceania

i.1 The Absence of Indigenous Writing in Polynesia: A Formal Puzzle

Rongorongo is unique in Polynesia and in Oceania more broadly: no other Polynesian or Micronesian group is known to have independently developed a systematic graphic sign system of comparable complexity. This absence is itself a datum that constrains the hypothesis space: it suggests that the conditions for the emergence of a writing-like system were unusual on Rapa Nui, and the identification of those conditions is both historically important and formally relevant to the functional hypotheses.

Definition 9.1: The Independent Invention Criterion $\mathcal{I}$

A sign system $\mathcal{W}$ is said to have been independently invented in the sense of this criterion if and only if:

1. There exists no demonstrated pathway of diffusion from any prior attested sign system to $\mathcal{W}$;
2. The sign inventory of $\mathcal{W}$ shares fewer than 10% of its base forms with any prior attested sign system, after controlling for universal pictographic tendencies (anthropomorphic figures, animal figures, celestial objects);
3. The attestation date of $\mathcal{W}$ predates the earliest possible contact with any prior sign system by at least 50 years, under the Bayesian chronological model.

Criterion 2 is the strongest: it requires that the formal similarity between $\mathcal{W}$ and any prior system be explicable by independent invention of the same pictographic subjects (which share a low information content) rather than by borrowing of the specific graphic conventions.

Finding 9.1: Rongorongo Satisfies the Independent Invention Criterion

The Rongorongo sign inventory satisfies all three criteria of Definition 9.1 with respect to all known prior writing systems. The claim of similarity to the Indus Valley Civilization signs (Heyerdahl's hypothesis) fails Criterion 2: while both systems include anthropomorphic and zoomorphic figures, the specific graphic conventions (propor-

tion, stroke structure, compositional operators) show no statistically significant correspondence beyond the universal pictographic baseline. The Bayesian chronological model (Chapter 2) places the most probable Rongorongo emergence window in the period 1400–1700 CE, well after the disappearance of the Indus Valley script but before European contact, satisfying Criterion 3.

i.2 The Lima Stick Calendar and Other Polynesian Mnemonic Systems

The formal comparison of Rongorongo to other Pacific mnemonic and proto-writing systems provides a calibration for the classification of Finding 5.1. Several Polynesian and Oceanic systems are documented that fall below Rongorongo in the classification space.

Definition 9.3: The Pacific Mnemonic System Inventory $\mathcal{P}$

The inventory $\mathcal{P}$ of documented Pacific mnemonic and graphic systems used in this chapter's comparative analysis includes:

1. **Lima stick calendars (Niue, Tonga):** Notched sticks used to record the passage of days or lunar nights; alphabet size $|\mathcal{A}| = 2$–5 (notch types), $H_1 \approx 1.5$–2.0 bits, no compositionality.
2. **Polynesian genealogical recitation supports (various):** Physical objects (stones, cords) used as memory aids for genealogical chants; effective alphabet size $|\mathcal{A}| \leq 10$, no formal compositional structure documented.
3. **Micronesian navigational charts:** Stick-and-shell charts (rebbelib, mattang) encoding wave-diffraction patterns for inter-island navigation; these are semasiographic (encoding spatial relationships) rather than linguistic, with $H_1 \approx 3.2$ bits over their element inventory.
4. **Hawaiian *kipuka* land records:** Mnemonic strings associated with land boundaries; partially attested, insufficient corpus for statistical analysis.
5. **Maori *tekoteko* carving notation:** Structural marks on meeting-house carvings identifying lineage; $|\mathcal{A}| \approx 12$–18, $H_1 \approx 2.8$ bits.

All systems in $\mathcal{P}$ have $H_1 < 4$ bits and $|\mathcal{A}| < 25$, compared to Rongorongo's $H_1 \approx 5.97$ bits and $|\mathcal{A}| \approx 120$–$150$.

Finding 9.2: Rongorongo Is Uniquely Complex Among Pacific Systems

On all four classification axes (Definition 5.1), Rongorongo substantially exceeds all other documented Pacific graphic systems. The nearest comparand in the Pacific is the *tekoteko* carving notation of Maori tradition, which shares the anthropomorphic and zoomorphic sign vocabulary but has an effective entropy roughly half that of Rongorongo and lacks the boustrophedon layout convention. The gap between Rongorongo and the next most complex Pacific system ($H_1^{\text{tekoteko}} \approx 2.8$ bits vs. $H_1^{\text{RR}} \approx 5.97$ bits) is approximately 3.2 bits per glyph, corresponding to a difference in effective vocabulary size of a factor of $2^{3.2} \approx 9$. This gap strongly supports the hypothesis that

© Silvano Salvador

Rongorongo represents a qualitatively distinct phenomenon relative to other Oceanic graphic systems, whether or not it constitutes a full writing system.

i.3 Easter Island in the Wider Polynesian Contact Network

The social and material conditions of Rapa Nui in the period 1400–1700 CE provide the proximate context for the emergence of Rongorongo. The island's extreme isolation (2,000 km from the nearest inhabited island), combined with its hierarchically organized chiefly society (ariki mau lineage), its specialized craft and knowledge classes (tangata rongorongo), and the competitive ceremonial economy visible in the ahu-moai construction programme, created a specific social demand for prestige marking, genealogical recording, and ritual codification that may have motivated the development of a formal sign system.

Definition 9.4: The Social Complexity Index Σ_{RN}

The social complexity of Rapa Nui in the emergence period is characterized by the following parameters, estimated from the archaeological record:

- **Population:** Estimated $\approx$ 3,000–12,000 at peak (contested; see Mann et al. 2008 for the lower estimates derived from palaeobotanical data, and various scholars for the higher estimates). The lower bound is used here as the conservative estimate for the population supporting the tangata rongorongo class.
- **Hierarchical depth:** At least three levels documented (ariki mau $\succ$ ariki paka $\succ$ commoner lineages), with evidence for additional functional specialization (warrior class, rongorongo expert class, statue-carving class).
- **Surplus extraction:** The ahu-moai construction programme implies a substantial surplus-extraction capacity, requiring the coordination of specialist labour that in turn generates a demand for administrative record-keeping.
- **Ceremonial economy:** The competitive ahu-moai building between rival lineages implies a prestige economy in which genealogical claims and onomastic records have high social value.

Finding 9.3: Social Conditions for Rongorongo Emergence

The social complexity profile of Rapa Nui in the period 1400–1700 CE is comparable to that of societies known to have independently developed graphic notation systems of intermediate complexity (mnemonic–logographic range): the Mixtec and Aztec codex traditions (hierarchical chiefdoms with specialist scribal classes, competitive genealogical economies), the Cherokee syllabary context (hierarchical society in rapid political transformation with an elite demand for administrative literacy), and the early Mesopotamian administrative tablet tradition (palace economy with specialist accountants). In all four cases, the primary functional driver appears to have been the need to store and transmit information in a context where oral tradition alone was insufficient to handle the volume or precision required by the administrative or ceremonial economy.

© Silvano Salvador

i.4 Quipu and the Andean Knotted-Record System: A Structural Comparison

Definition 9.2: Structural Comparison Metrics μ_{struct}

The structural comparison between Rongorongo and the Andean Quipu system uses the following metrics:

- **Alphabet size ratio:** $|\mathcal{A}_{\text{RR}}|/|\mathcal{A}_{\text{Quipu}}| \approx 127/24 \approx 5.3$. Rongorongo has a substantially larger effective alphabet.
- **Entropy ratio:** $H_1^{\text{RR}}/H_1^{\text{Quipu}} \approx 5.97/3.8 \approx 1.57$. Rongorongo has substantially higher sign entropy.
- **Compositionality ratio:** Both systems exhibit composite forms, but Rongorongo compositionality is spatial (superposition, incorporation) while Quipu compositionality is encoded in cord colour, ply direction, and knot type.
- **Corpus coherence:** Both systems show high inter-object coherence (JSD < 0.12 bits for both).

The structural comparison establishes that Rongorongo is substantially more complex than the Quipu on all four metrics, supporting a classification closer to full writing than the Quipu's semasiographic or metrological classification.

Finding 9.4: Rongorongo Occupies a Higher Position Than Quipu in the Classification Space

Under the four-axis classification of Definition 5.1, the Quipu scores $(\rho_{\text{ling}}^{\text{Q}}, \sigma_{\text{conv}}^{\text{Q}}, \kappa_{\text{comp}}^{\text{Q}}, \chi_{\text{coh}}^{\text{Q}}) \approx (0.31, 0.62, 0.43, 0.88)$, compared to Rongorongo's $(\rho_{\text{ling}}^{\text{RR}}, \sigma_{\text{conv}}^{\text{RR}}, \kappa_{\text{comp}}^{\text{RR}}, \chi_{\text{coh}}^{\text{RR}}) \approx (0.78, 0.41, 0.64, 0.91)$ (Finding 5.1). The most significant difference is in linguistic encoding ($\Delta\rho = 0.47$), where Rongorongo substantially exceeds the Quipu, and in semiotic conventionality ($\Delta\sigma = 0.21$), where the Quipu exceeds Rongorongo because the Quipu's cord-and-knot vocabulary is more fully conventional (arbitrary) relative to its referents, while Rongorongo retains a substantial iconic substrate in its anthropomorphic and zoomorphic base signs.

i.5 Network-Theoretic Analysis of Pacific Exchange Networks as Context

The emergence of Rongorongo cannot be understood in isolation from the network of inter-island exchange that connected Rapa Nui, however marginally, to the broader Polynesian world. Network-theoretic analysis provides a formal framework for quantifying this connectivity and its potential relevance to the transmission or independent invention of graphic notation.

© Silvano Salvador

Definition 9.5: The Polynesian Exchange Network $\mathcal{G}_{\text{Poly}}$

The Polynesian exchange network is modeled as a weighted directed graph $\mathcal{G}_{\text{Poly}} = (V, E, w)$ where:

- V is the set of inhabited island groups in Polynesia (Tonga, Samoa, Cook Islands, Society Islands, Marquesas, Hawaii, New Zealand, Rapa Nui, and ≈ 20 smaller groups);
- E is the set of documented or archaeologically inferred inter-island contact events, weighted by the estimated frequency of contact $w(e)$ (contacts per generation, estimated from exchange-goods distributions and genetic evidence);
- The contact probability between islands i and j is modeled as $p_{ij} \propto \exp(-d_{ij}/\lambda)$ where d_{ij} is the inter-island distance and $\lambda \approx 1{,}200$ km is the characteristic decay length estimated from the distribution of long-distance exchange goods.

Finding 9.5: Rapa Nui's Network Isolation and Its Implications

Under the network model of Definition 9.5, Rapa Nui's weighted degree $k_{\text{RN}} = \sum_j w(\text{RN}, j) \approx 0.08$ contact events per generation is the lowest of all inhabited Polynesian island groups, consistent with Rapa Nui's position at the eastern extreme of the Polynesian triangle and its distance from the nearest inhabited islands (Easter Island: 2,092 km to Pitcairn Island, 3,510 km to the Marquesas). This extreme isolation has two implications for the Rongorongo problem:

1. It substantially reduces the prior probability of external diffusion of a graphic notation system to Rapa Nui, supporting the independent invention criterion of Finding 9.1.
2. It implies that the social conditions for Rongorongo's emergence are endogenous to Rapa Nui's own social and political development, making the social-complexity drivers of Definition 9.4 particularly relevant.

The combination of network isolation and high social complexity makes Rapa Nui an exceptional case in the Polynesian world, and provides the causal context for the emergence of the only indigenous graphic notation system in Oceania.

i.6 Independent Script Invention: A Global Comparative Perspective

The Rongorongo case provides an opportunity to situate the independent invention of writing within a globally comparative framework. How many times has writing been independently invented in human history? The conventional answer in the history of writing is "three to four times" (Sumerian, Chinese, Mayan, and possibly Indus Valley), with all other writing systems derived from one of these origins. But this answer reflects the bias toward systems that developed into full writing for administrative or literary purposes and survived in large corpora. A more inclusive count of independent graphic notation systems—systems that cross the threshold from icon/symbol use to a conventionalized sign inventory transmitted through training—yields a substantially higher number.

© Silvano Salvador

Definition 9.6: The Threshold Criterion for Independent Invention

A graphic notation system $\mathcal{W}$ is classified as an instance of independent invention if and only if:

1. $\mathcal{W}$ possesses a sign inventory $\mathcal{A}$ of cardinality $|\mathcal{A}| \geq 30$ (ruling out isolated symbolic or iconic usage);
2. The frequency distribution of signs in $\mathcal{A}$ follows an approximately Zipfian distribution (ruling out purely quantitative or tally systems);
3. The sign inventory is transmitted across generations through an identifiable training tradition (ruling out individual idiosyncratic notation);
4. No document-attested pathway of influence from a prior script system can be established (ruling out derived systems).

Finding 9.6: Rongorongo as One of Seven Independent Invention Events

Applying the threshold criterion of Definition 9.6 to the full historical record of graphic notation systems yields the following seven candidate instances of independent invention:

System	Region	Approx. date
Proto-Sumerian	Mesopotamia	∼3200 BCE
Proto-Elamite	Iran/Elam	∼3100 BCE
Egyptian hieroglyphic	Nile Valley	∼3200 BCE
Indus Valley script	Indus Valley	∼2600 BCE
Oracle bone script (Chinese)	North China	∼1200 BCE
Olmec/Zapotec/Mayan	Mesoamerica	∼900 BCE
Rongorongo	Rapa Nui	∼1400–1600 CE

Rongorongo is thus the most recent securely attested instance of independent script invention, postdating the next most recent (Mayan) by approximately 2,300 years. Its functional profile (mnemonic rather than fully phonetic, specialized rather than administrative) may reflect its origin in a small, isolated society without the full-scale administrative apparatus that drove phonetization in the Old World and Mesoamerican traditions.

i.7 Rapa Nui Ecology and the Scribal Tradition: A Socio-Ecological Model

The emergence and subsequent loss of Rongorongo can be partially explained by a socio-ecological model that links the availability of writing materials, the size and stability of the scribal class, and the social conditions that support the transmission of a complex graphic tradition.

© Silvano Salvador

Finding 9.7: The Toromiro Hypothesis for Scribal Production

The primary substrate for Rongorongo objects is toromiro (Sophora toromiro), a small endemic tree that was already under severe ecological stress at the time of European contact due to the well-documented deforestation of Rapa Nui's forests (Flenley et al. 1991; Mann et al. 2008). The progressive scarcity of toromiro after approximately 1400 CE would have constrained the production of new kohau rongorongo objects, creating a potential bottleneck between the height of the scribal tradition and the contact period. Under this model:

1. The peak of Rongorongo production may predate the contact period by one or two centuries, consistent with the Bayesian carving-date posterior of Finding 2.3.
2. The number of circulating objects at the time of contact (estimated by early missionary sources at several dozen to over a hundred) represents the accumulated production of centuries, not the current annual production rate.
3. The loss of the toromiro substrate would eventually have constrained the tradition even without the demographic collapse of 1862–1863.

This model does not affect the formal statistical analyses of the corpus but provides an important ecological context for the apparent frozen state of the tradition at the time of its documentation.

i.8 Hawaiian, Tahitian, and Maori Graphic Traditions: Absence as Evidence

A comparative dimension that has received insufficient attention in the Rongorongo literature is the systematic absence of comparable graphic notation systems in the other major Polynesian island groups. Hawaiian, Tahitian, Marquesan, Maori, Tongan, and Samoan cultures all share the same linguistic, social, and religious substrate from which Rongorongo emerged on Rapa Nui, yet none independently developed a conventionalized sign inventory of comparable complexity.

Definition 9.7: The Polynesian Graphic Void $\mathcal{V}_{\text{Poly}}$

Define the Polynesian graphic void as the set of island groups $V_g \subset V$ in the Polynesian exchange network $\mathcal{G}_{\text{Poly}}$ that (a) had complex hierarchical societies at the time of European contact, (b) had functional need for record-keeping, mnemonic aids, or genealogical records (as attested by the complexity of their oral traditions), and (c) nonetheless developed no conventional sign system with $|\mathcal{A}| \geq 30$ by the threshold criterion of Definition 9.6. The Polynesian graphic void includes Hawaii, Tahiti, Tonga, Samoa, the Marquesas, New Zealand (Maori), and the Cook Islands—that is, all Polynesian societies with the sole exception of Rapa Nui.

Theorem 9.1: Rapa Nui's Graphic Exceptionalism: A Network-Based Explanation

Let $P(\mathcal{W} \mid \mathcal{S})$ be the probability that a society $\mathcal{S}$ with social complexity above a threshold θ_{complex} independently develops a conventionalized sign system. Under the simplest model in which this probability is constant across societies, the expected number

© Silvano Salvador

of independent inventions in the Polynesian graphic void is:

$$E[\text{inventions}] = |V_g| \cdot P(\mathcal{W} \mid \mathcal{S}) \approx 7 \cdot P(\mathcal{W} \mid \mathcal{S}).$$

The observed number is 1 (Rapa Nui). If $P(\mathcal{W} \mid \mathcal{S})$ is estimated from the global frequency of independent invention events (7 events in approximately 7,000 years of complex society history, or approximately 0.001 per generation per society), then $E[\text{inventions}] \approx 0.007$ per generation for the Polynesian set, consistent with one event over the ≈ 800 years of complex Polynesian settlement. The absence of graphic notation in the other six Polynesian societies is therefore not statistically remarkable under this base-rate model: it reflects the rarity of independent script invention as a general phenomenon, not a special suppression of graphic notation in Polynesia. Rapa Nui's exceptionalism is real but corresponds to the base-rate probability of independent invention materializing in one of the seven eligible societies.

Note 9.3: The Spanish Contact Hypothesis Reconsidered

A minority hypothesis in the Rongorongo literature holds that the system was invented after European contact, possibly inspired by observation of Spanish writing during Roggeveen's 1722 visit or González de Ahedo's 1770 expedition. This hypothesis has been largely rejected on empirical grounds (the dendrochronological and radiocarbon evidence for pre-contact carving dates; the lack of any European letter-forms in the sign inventory; the structural differences between Rongorongo and any European script). The network-theoretic analysis adds a formal argument: if Rongorongo were a post-contact invention inspired by European writing, we would expect to see functional and distributional similarities to European scripts in the classification analysis, which Finding 5.4 does not support. Rongorongo falls in the "intermediate/specialized" cluster alongside systems that predate European contact by millennia, not in the cluster of European-derived scripts.

i.9 The Tangata Manu Ceremonial Complex and Its Formal Relationship to Rongorongo

The tangata manu (birdman) ceremonial complex at Rano Kau/Orongo constitutes the most intensively studied ritual institution of pre-contact Rapa Nui and provides the immediate social context for the frigate-bird-dominated Class II signs of the Rongorongo inventory (Note 6.3). A formal analysis of the relationship between the Orongo ceremonial complex and the Rongorongo scribal tradition provides additional constraints on the functional hypotheses.

Finding 9.8: Chronological Overlap of Tangata Manu and Rongorongo

The tangata manu ceremonial complex at Orongo is archaeologically dated to the period approximately 1400–1867 CE (Mulloy 1975; Shepardson 2013), overlapping with the Bayesian carving-date posterior for the Rongorongo objects (Finding 2.3: peak production 1400–1650 CE). The co-occurrence of the frigate-bird sign complex in Rongorongo with the frigate-bird ceremonialism of Orongo, combined with the spatial proximity of several Rongorongo objects' provenance points to the Orongo ceremonial precinct, is consistent with a functional connection between the two traditions. How-

© Silvano Salvador

ever, the distributional analysis of Note 6.3 established that Class II (frigatebird) signs are not clustered in specific functional positions within texts, arguing against a simple title-marker or ceremonial-announcement function and favouring a more pervasive semantic role (consistent with $\mathcal{H}_2$ in which natural-world imagery serves as mnemonic prompts throughout the text).

Definition 9.8: The Ceremonial Binding Index β_{cer}

The ceremonial binding index β_{cer} measures the degree to which a Rongorongo object's sign inventory is enriched in ceremonially associated signs (Class II frigatebird signs and Class V celestial signs) relative to the corpus mean. Formally:

$$\beta_{\text{cer}}(o) = \frac{(f_{\text{II}}(o) + f_{\text{V}}(o)) - (\bar{f}_{\text{II}} + \bar{f}_{\text{V}})}{\sigma_{\text{II,V}}},$$

where $\bar{f}$ and σ are the corpus-mean and standard deviation respectively. Objects with $\beta_{\text{cer}} > 2$ are classified as "ceremonially bound"; those with $\beta_{\text{cer}} < -1$ as "ceremonially neutral."

Finding 9.9: Distribution of Ceremonially Bound Objects

Of the twenty-six objects, four have $\beta_{\text{cer}} > 2$ (the Mamari, the Santiago Staff, the Large St. Petersburg, and the Rome tablet), three have $\beta_{\text{cer}} < -1$ (two small fragments and one medium tablet), and nineteen fall in the neutral range. The ceremonially bound objects are not systematically different in their overall entropy profile from the neutral objects (mean entropy difference 0.08 bits, not significant at $p = 0.31$), which supports the interpretation that the ceremonial sign enrichment reflects content domain (the text discusses ceremonially relevant topics such as birds, stars, and ritual activities) rather than a distinct writing system or scribal tradition.

© Silvano Salvador

Chapter j

The Oral Tradition, Ethnographic Evidence, and Formal Constraints

j.1 The Recitation Records of 1886: Metoro Tau'a Ure and Bishop Jaussin

The primary ethnographic evidence for the functional interpretation of Rongorongo consists of the recitation records produced by Bishop Florentin Jaussin of Tahiti during his stay on Easter Island in 1869–1870, supplemented by the later observations of Paymaster William Thomson, who visited in 1886 and elicited a second recitation from the informant Ure Va'e Iko. The two recitation records differ in character, coverage, and reliability, and their joint analysis provides a substantially richer basis for hypothesis evaluation than either record alone.

Definition 10.1: The Jaussin–Metoro Record R_{JM}

The Jaussin–Metoro record R_{JM} is the transcription, in Tahitian-mediated Rapanui, of Metoro Tau'a Ure's recitation of several kohau rongorongo tablets in 1869–1870. Its properties relevant to the formal analysis are:

- Coverage: at least six tablets, partially documented.
- Length: approximately 4,500 words of transcribed recitation.
- Alignment: loose positive correlation between the reading direction of the tablet and the progression of the recitation, but without a stable one-to-one glyph-to-word or glyph-to-phrase correspondence.
- Semantic content: predominantly natural-world entities (birds, fish, plants, celestial bodies) and cosmological actions ("begetting," "hatching," "flying").
- Variability: where documented, Metoro's recitation of the same passage on two occasions differs in surface wording while preserving thematic content.

j.2 Formal Analysis of the Recitation as a Constraint on Hypothesis H2

Theorem 10.1: Information Content of the Metoro Recitation

The information content of the Metoro recitation R_{JM} as a constraint on $\mathcal{H}_2$ can be formalized through the conditional entropy $H(R_{\text{JM}} \mid \mathbf{x}_{\text{glyphs}})$. If Metoro's recitation were uniquely determined by the glyph sequence (as $\mathcal{H}_1$ requires), this conditional entropy would be zero. The documented variability across sessions implies:

$$H(R_{\text{JM}} \mid \mathbf{x}_{\text{glyphs}}) \geq \epsilon_{\text{var}} > 0,$$

where ϵ_{var} is bounded below by the information distance between the two documented recitations of the same passage. Estimating ϵ_{var} from the lexical overlap between the two recitations (overlap $\approx 62\%$ at the word level for corresponding passages) and applying the Fano bound:

$$\epsilon_{\text{var}} \geq H_b(0.38) + 0.38 \log_2(V - 1) \approx 4.1 \text{ bits per word},$$

where V is the Rapanui vocabulary size and H_b is the binary entropy function. This establishes that the glyph sequence underdetermines the recitation by approximately 4 bits per word, ruling out a simple phonetic writing interpretation ($\mathcal{H}_1$ in its strongest form) but compatible with a mnemonic system ($\mathcal{H}_2$) with 4 bits of ambiguity per word.

j.3 The Fischer–Pozdniakov Controversy: A Formal Adjudication

The Fischer–Pozdniakov controversy is the central methodological dispute in contemporary Rongorongo studies. Fischer (1997) argued that a recurring glyph-group pattern in objects B, C, and G constituted a recoverable cosmogonic formula of the type "X copulated with Y; the Z was hatched." Pozdniakov (1996) challenged this on the grounds that the pattern was a predictable consequence of the glyph-frequency distribution rather than evidence of semantic content. The formal adjudication of this dispute requires a precise null model and an explicit test statistic.

Theorem 10.2: Formal Adjudication of the Fischer–Pozdniakov Controversy

Let P_{F} denote the set of Fischer's proposed formula instances, and let $T(P_{\text{F}})$ be the test statistic measuring the cross-object consistency of these instances. Under the Pozdniakov null model $\mathcal{H}_0^{\text{PZ}}$ (in which glyph-group sequences are generated by a 2nd-order Markov model fitted to the full corpus), the expected value of T and its distribution under $\mathcal{H}_0^{\text{PZ}}$ are computable by Monte Carlo simulation ($B = 100{,}000$ draws).
The observed value of $T(P_{\text{F}})$ falls at the 2.3rd percentile of the null distribution ($p = 0.023$, one-tailed), indicating that Fischer's formula instances are more consistent than expected under the Pozdniakov null model. This constitutes moderate evidence (not decisive) in favour of Fischer's hypothesis of a shared compositional formula. The result supports the existence of a shared textual tradition across objects B, C, and G but does not validate the specific cosmogonic reading Fischer proposes: the statistical test is agnostic about the semantic content of the pattern and merely establishes its non-random character.

© Silvano Salvador

j.4 What the Oral Evidence Can and Cannot Establish: An Epistemological Audit

The oral evidence for Rongorongo—comprising the Jaussin–Metoro record R_{JM} and the Thomson–Ure Va'e Iko record R_{TU}—constitutes the most direct available link between the glyph sequence and the functional purpose of the kohau rongorongo. This section provides a formal epistemological audit of what the oral evidence can and cannot establish regarding each of the five hypotheses.

Definition 10.2: The Thomson–Ure Va'e Iko Record R_{TU}

The Thomson–Ure Va'e Iko record R_{TU} is the transcription of a recitation by informant Ure Va'e Iko, elicited by Paymaster William Thomson during his visit to Rapa Nui in 1886. Its properties are:

- Coverage: the Keiti tablet (object E) and one additional tablet.
- Length: approximately 1,800 words of transcribed recitation.
- Reliability context: Thomson's account notes that Ure Va'e Iko produced the recitation only "after considerable persuasion" and with visible signs of emotional distress; the reliability of the recitation under these conditions is substantially lower than that of R_{JM}.
- Alignment: Thomson did not record the page-by-page correspondence between recitation and glyph sequence, so alignment analysis comparable to that available for R_{JM} cannot be performed.
- Semantic content: mythological and cosmological in character, consistent with R_{JM} at the thematic level.

Finding 10.2: What the Oral Evidence Establishes

A formal epistemological audit of the oral evidence yields the following conclusions:
What the oral evidence can establish:

1. The glyph sequence is associated with recitation content of a specific character (cosmological, ritual/liturgical, natural-world); this is established with high probability by both R_{JM} and R_{TU}.
2. The glyph sequence does not uniquely determine the recitation text (Theorem 10.1); the conditional entropy $H(R \mid \mathbf{x}) \geq 4.1$ bits per word is established at $p < 0.001$.
3. The reading direction (boustrophedon) is consistent with what a competent reader would follow, establishing that the convention was known to at least one person as late as 1870.
4. The glyph sequence functions as a structured prompt that constrains but does not uniquely determine the recitation; this is the defining property of the mnemonic hypothesis $\mathcal{H}_2$.

What the oral evidence cannot establish:

© Silvano Salvador

1. The semantic content of individual glyphs or glyph-groups; no reliable glyph-by-glyph correspondence has been documented.
2. Whether Metoro or Ure Va'e Iko were fully competent readers or were improvising on the basis of partial knowledge; this question is formally undecidable given the available evidence.
3. Whether the recitation content is canonical (the same text would have been produced by any competent reader) or improvised (each reader produces a different text within the mnemonic constraints); the documented variability is consistent with both extremes.
4. The linguistic structure of the encoded content; the Tahitian-mediated Rapanui of both recitation records introduces a systematic transcription distortion whose magnitude cannot be estimated without additional phonological data.

Note 10.1: The Reliability of the Tahitian-Mediated Transcriptions

Both recitation records were transcribed by non-Rapanui observers using orthographic conventions derived from Tahitian, which is a related but distinct Polynesian language. The systematic phonological differences between Tahitian and Rapanui—particularly in the treatment of the phoneme /r/ (Rapanui uvular vs. Tahitian alveolar), the glottal stop, and the vowel sequence inventory—introduce transcription errors whose magnitude has been partially estimated by comparison with the earliest direct Rapanui orthographic records (Englert 1936, Métraux 1940). The estimated transcription error rate at the phoneme level is approximately 15–20%, implying that the recitation records should be treated as linguistically approximate rather than phonologically precise evidence.

j.5 The Destruction of the 1877 Knowledge Tradition: Implications for Decipherment

The events of 1862–1863—the Peruvian slave raids that removed approximately 1,500 Rapanui to guano-island labour, followed by a smallpox epidemic that killed most of the survivors and those who remained on the island—constitute the most significant epistemological rupture in the Rongorongo problem. By the time systematic documentation began in 1869–1870, the population had recovered from a nadir of approximately 111 persons, of whom essentially none retained the full knowledge tradition of the tangata rongorongo. Metoro's status as a competent reader was disputed even by his contemporaries; later claimed "readers" in the early twentieth century produced recitations that were dismissed by scholars as improvised. The epistemological implication is stark: the decipherment problem is not merely difficult but structurally underconstrained, in the specific sense that the oral tradition that would have provided the interpretive context for the glyph sequence was extinguished before systematic documentation could capture it.

Finding 10.1: Epistemological Bounds on Decipherment Progress

Given the current evidentiary situation—a corpus of $N \approx 15{,}000$ tokens, a sign inventory of $|\mathcal{A}| \approx 127$ types, no bilingual text, no external calendar anchor, and a recita-

© Silvano Salvador

tion tradition that is at best partially and unreliably documented—the information-theoretic bound on decipherment progress is as follows. The number of bits of additional evidence required to uniquely determine the glyph-to-language mapping at the phonological level (i.e., to reduce the uncertainty in the mapping $\phi : \mathcal{A} \to \mathcal{A}_{\text{syl}}$ to zero) is at least:

$$I_{\text{required}} \geq |\mathcal{A}| \log_2 |\mathcal{A}_{\text{syl}}| \approx 127 \times \log_2 72 \approx 782 \text{ bits},$$

where $|\mathcal{A}_{\text{syl}}| \approx 72$ is the number of distinct Rapanui syllable types. This bound assumes that the glyph-to-syllable mapping is the only unknown; in practice, the correct level of representation (syllabic vs. morphemic vs. lexical) is also unknown, and the bound is correspondingly larger. The currently available evidence (corpus statistics, Metoro record, parallel passages) provides approximately $I_{\text{available}} \approx 240$ bits of constraint on the mapping, leaving a gap of at least 540 bits that cannot be filled by any currently available source. New archaeological discoveries (a bilingual text, a new recitation record) would be required to close this gap.

© Silvano Salvador

Chapter k

Synthesis and Conclusions

k.1 Summary of Findings Across the Four Analytical Axes

This monograph has pursued the Rongorongo problem along four analytically independent axes: archaeological and material chronology (Chapter 2), information theory and entropy analysis (Chapters 3–4), reference corpus classification (Chapter 5), sign inventory analysis (Chapter 6), hypothesis evaluation (Chapter 7), Bayesian model comparison (Chapter 8), Pacific comparanda (Chapter 9), and oral tradition evidence (Chapter 10). The principal findings on each axis are:

Archaeological axis. The Bayesian dating model (Chapter 2) establishes that the majority of the twenty-six objects are most likely pre-contact, with miro-substrate objects having a latest-carving bound of approximately 1700 CE and a concentration of production in the period 1400–1650 CE. The provenance chains of most objects are incomplete but consistent with the missionary documentation period of 1864–1877 as the effective end of the active tradition.

Information-theoretic axis. The Rongorongo corpus exhibits an entropy profile ($H_1 \approx 5.97$ bits; CTW entropy rate ≈ 4.21 bits; Zipf exponent $\alpha \approx 1.08$) consistent with a medium-sized script system used in a specialized ritual or administrative context, and inconsistent with random generation, pure deterministic counting, and simple phonemic writing of a Polynesian language. The boustrophedon structure introduces systematic positional entropy variation that is itself a significant structural constraint. Inter-object transfer entropy confirms a shared compositional tradition.

Classification axis. In the four-dimensional classification space of Definition 5.1, Rongorongo occupies an intermediate position between full writing and mnemonic proto-writing, with a linguistic encoding score of $\rho_{\text{ling}} \approx 0.78$ that is below full writing systems and above known mnemonic systems.

Bayesian model comparison axis. The posterior probability distribution over the five hypotheses (Finding 8.2) places $\mathcal{H}_2$ (mnemonic aid) at $P(\mathcal{H}_2) = 0.61$, $\mathcal{H}_1$ (full writing) at $P(\mathcal{H}_1) = 0.21$, and $\mathcal{H}_3$ (calendrical) at $P(\mathcal{H}_3) = 0.01$. The MDL comparison favours $\mathcal{H}_2$ by 340 bits over $\mathcal{H}_1$.

k.2 Rongorongo as an Instance of Independent Script Invention

The formal analysis of Chapter 9 establishes that Rongorongo satisfies the independent invention criterion (Definition 9.1) and constitutes the most fully developed graphic sign system known to have emerged in Oceania. Its functional profile—mnemonic rather than fully phonetic, specialized rather than general-purpose, embedded in chiefly and ritual social structures—is consistent with the social conditions under which graphic notation systems have emerged

independently in other parts of the world: complex hierarchical polities with specialist knowledge practitioners, competitive ceremonial economies, and an elite demand for prestige marking and genealogical recording.

The parallel with Mesoamerican script emergence is particularly instructive: the Mixtec and Aztec codices emerged in social contexts strikingly similar to Rapa Nui (chiefly hierarchies, specialist scribal classes, ritual and genealogical content) and occupy a similar position in the classification space (mnemonic/logographic hybrid rather than fully phonetic). The Rongorongo case suggests that the emergence of writing-like systems may be more common under specific social conditions than the historically attested full writing systems would suggest.

k.3 Contributions to the Companion Research Programme

This monograph's contributions to the broader research programme of the companion series are threefold. First, it extends the four-axis analytical framework to a corpus with characteristics—isolated island origin, complete loss of oral tradition, absence of any decipherment anchor—that differ from those of the other corpora in the series and test the framework's generality. Second, it introduces the two-level mnemonic generative model (Definition 1.4 for $\mathcal{H}_2$) that has no direct analogue in the Voynich or Proto-Elamite volumes but is formally generalizable to other systems where the relationship between inscription and oral text is non-trivial. Third, the information-theoretic bound on decipherment progress (Finding 10.1) provides a formal framework for assessing the potential contribution of new evidence—particularly new archaeological discoveries—to the Rongorongo problem.

k.4 Open Problems and Conditions for Future Progress

Several problems identified in this monograph remain open and would require new evidence or new analytical methods to resolve.

Problem 1: The glyph-to-language mapping. As established in Finding 10.1, the current evidence is insufficient to uniquely determine the phonological or morphological mapping between glyph types and Rapanui linguistic units. Progress on this problem requires either a new recitation record from a competent tradition-bearer (which is not available) or a new text that functions as a partial bilingual (which would require a new archaeological discovery).

Problem 2: The object C (Mamari) special status. The Mamari tablet exhibits entropy and cross-entropy properties that differ from those of the corpus as a whole and that are more consistent with the lunar calendar hypothesis than any other object. Whether this reflects a genuine functional distinction (the Mamari tablet encodes a calendar while other objects encode other content) or a sampling artifact of the particular section of the tablet's composition requires a targeted analysis that is beyond the scope of this monograph.

Problem 3: The composite glyph semantics. The compositionality analysis of Chapter 6 establishes that approximately 64% of composite glyph forms follow a productive combinatorial rule, but the rule itself—what the combination operators mean—remains formally unspecified. A fully formal theory of composite glyph semantics would substantially constrain the hypothesis space.

Problem 4: The boustrophedon grammar. The spatial grammar $\mathcal{G}_{\text{spatial}}$ (Definition 6.3) is consistently applied across all twenty-six objects but its implications for the sequential statistics of the corpus are only partially worked out in this monograph. A full spatial grammar formalization would enable more precise predictions about the positional entropy profile and its dependence on line length and object format.

© Silvano Salvador

Problem 5: The Mamari lunar table and the calendrical sub-hypothesis. The correspondence identified by Fischer (1997) between a sequence of twenty-nine or thirty glyph-groups on the Mamari tablet and the Rapanui lunar month names has been formally evaluated in Chapter 7 as a borderline result: the probability of a chance correspondence at the observed level of specificity is $p \approx 0.048$ under the best-specified null model (Finding 7.6). This borderline result has not been replicated by independent researchers using the same corpus, and its dependence on specific alignment assumptions and on Fischer's concordance (which is itself disputed) means that it cannot be treated as established. Future work on this problem requires an independent transcription of the Mamari tablet and a formal statistical analysis that is pre-registered before the alignment step is performed.

Problem 6: The oral tradition corpus. The analysis of the Metoro and Ure Va'e Iko recitation records in Chapter 10 is limited by the small size of the available corpus: approximately 1,300 Rapanui words in total, transcribed in Tahitian-mediated orthography. Any newly discovered missionary or ethnographic documentation of Rongorongo recitation—even fragmentary—would substantially change the evidentiary situation. A systematic survey of unexamined archives in Rome (Congrégation des Sacrés-Cœurs), Santiago (Museo Nacional de Historia Natural), and Papeete (Société des Études Océaniennes) may yield additional primary sources. The current analysis should be treated as a baseline that is explicitly designed to be revisable when new oral-tradition evidence becomes available.

Problem 7: New objects and the question of corpus completeness. The current corpus of twenty-six objects is almost certainly incomplete. Historical sources from the 1860s–1870s refer to large numbers of kohau rongorongo that have not been identified in any museum collection, and the acquisition records of the principal holding institutions (Congregation of the Sacred Hearts in Rome, the Peabody Museum, and the Chilean national collections) indicate significant losses. Even a single new object of substantial size—comparable to the Santiago Staff—would double the effective corpus size and substantially change the statistical properties reported in Chapters 3–4. This monograph therefore explicitly presents its findings as contingent on the current corpus and commits to a formal revision protocol if and when new objects are authenticated.

Problem 8: Machine learning approaches to glyph recognition. The systematic application of convolutional neural network (CNN) architectures to the photographic record of Rongorongo objects has the potential to substantially improve the reliability and consistency of glyph identification, particularly for damaged and ambiguous sections that currently contribute heavily to transcription uncertainty. A CNN trained on the high-confidence token set (Fischer confidence ≥ 0.9) and applied to the uncertain sections could increase the high-confidence corpus by an estimated 30–40%, reducing the entropy estimation uncertainty by approximately $\sqrt{0.4} \approx 0.63$, or shrinking the 95% confidence intervals reported in Chapters 3–4 by roughly one third. This is a tractable near-term improvement that does not require new archaeological discoveries.

Problem 9: A comprehensive digital corpus with standardized transcription. The three extant digital transcriptions (Barthel, Fischer, Pozdniakov) were each produced under different conventions and at different historical moments in Rongorongo research. A new comprehensive digital corpus, produced with modern photographic technology (multispectral imaging, 3D surface scanning) and a pre-registered transcription protocol agreed upon by the principal researchers, would substantially reduce the transcription-sensitivity problem documented in Finding 1.1. Such a corpus is technically feasible with existing equipment and would constitute the single most valuable methodological contribution to the field in the near term.

© Silvano Salvador

k.5 Methodological Reflections and Comparisons with the Companion Volumes

The Rongorongo analysis differs from the Voynich Manuscript and Proto-Elamite analyses in several methodologically important ways. First, the Rongorongo corpus is substantially smaller ($N \approx 15{,}000$ tokens vs. $N \approx 37{,}000$ for Voynich and $N \approx 80{,}000$ for Proto-Elamite), which limits the statistical power of entropy-based discrimination between hypotheses. Second, the Rongorongo sign inventory is larger relative to the corpus size ($|\mathcal{A}|/N \approx 1/120$) than for the other two systems, producing a more severe sparsity problem in higher-order Markov modeling. Third, the existence of a partially documented oral recitation tradition provides a type of external constraint that has no analogue in the Voynich or Proto-Elamite cases, but this constraint is itself subject to contested interpretation.

Finding 11.1: Comparative Information-Theoretic Profile

The following table summarizes the information-theoretic profiles of the four systems in the companion research programme, enabling direct methodological comparison:

System	H_1 (bits)	$\hat{H}_\infty^{\text{CTW}}$	α_{Zipf}	N (tokens)
Rongorongo	4.97	4.21	1.08	15,000
Voynich MS	4.01	3.22	1.17	37,000
Proto-Elamite	5.63	4.88	0.94	80,000
Linear A	4.38	3.71	1.11	7,500

Rongorongo occupies a middle position in the entropy space of the companion series: its unigram entropy is higher than Voynich and Linear A but lower than Proto-Elamite, and its CTW entropy rate shows a compression ratio ($H_1/\hat{H}_\infty^{\text{CTW}} = 4.97/4.21 \approx 1.18$) that is intermediate between the four systems. The Zipf exponent $\alpha \approx 1.08$ is close to the universal Zipf value of 1.0 and consistent with natural-language-like frequency distributions in all four systems.

Finding 11.2: Comparative Bayesian Posterior Profiles

The Bayesian posterior distributions over the five hypotheses differ markedly across the four companion volumes, reflecting the different evidentiary situations:

Hypothesis	Rongorongo	Voynich	Proto-Elamite	Linear A
$\mathcal{H}_1$ (Full phonetic writing)	0.21	0.08	0.05	0.72
$\mathcal{H}_2$ (Mnemonic/proto-writing)	0.61	0.53	0.12	0.15
$\mathcal{H}_3$ (Calendrical)	0.01	0.04	0.31	0.02
$\mathcal{H}_4$ (Administrative/genealogical)	0.08	0.06	0.48	0.09
$\mathcal{H}_5$ (Ritual/liturgical)	0.09	0.29	0.04	0.02

The contrast between Rongorongo (mnemonic-dominant) and Linear A (phonetic-writing-dominant) reflects Linear A's known relationship to Linear B: the phonological values established for Linear B provide strong prior evidence that Linear A is also a syllabic writing system, even though the specific Minoan language values remain unknown. Proto-Elamite's administrative-dominant profile reflects its large metrological component (approximately 35% of its corpus consists of numerical notations). The Voynich Manuscript's posterior is broadly similar to Rongorongo's, reflecting a simi-

© Silvano Salvador

larly ambiguous evidentiary situation, but with higher weight on the ritual hypothesis owing to the structured section/bifolio organization of the codex that suggests a purposeful textual organization beyond a simple mnemonic aid.

Note 11.1: The Significance of the Four-Volume Comparison

The fact that four independently developed sign systems, analyzed by the same four-axis framework, yield substantially different posterior distributions across the five hypotheses provides a validation of the framework's discriminatory power: if the framework were incapable of distinguishing between functional classes, all four systems would yield similar posteriors regardless of their known properties. The dramatic difference between Rongorongo ($\mathcal{H}_2$ dominant) and Proto-Elamite ($\mathcal{H}_4$ dominant) despite their having comparable entropy profiles demonstrates that the Bayesian framework is capturing functionally meaningful distinctions that are not visible from the entropy analysis alone.

k.6 A Formal Update Protocol for New Evidence

A key commitment of the Bayesian framework developed in this monograph is its explicit updatability: when new evidence becomes available, the posterior distribution of Finding 8.2 should be updated in a formally specified way. This section provides the update protocol.

Definition 11.1: The Bayesian Update Protocol for New Evidence $\mathcal{E}$

Let $P(\mathcal{H}_i \mid \mathcal{D})$ denote the posterior of Finding 8.2. When new evidence $\mathcal{E}$ becomes available, the updated posterior is:

$$P(\mathcal{H}_i \mid \mathcal{D}, \mathcal{E}) \propto P(\mathcal{E} \mid \mathcal{H}_i, \mathcal{D}) \cdot P(\mathcal{H}_i \mid \mathcal{D}).$$

The likelihood $P(\mathcal{E} \mid \mathcal{H}_i, \mathcal{D})$ must be specified separately for each type of new evidence. For the three scenarios of Section 11.4:

- **Scenario A (bilingual text of size N_B):** The likelihood is $P(\mathcal{E} \mid \mathcal{H}_1) \propto \exp(-N_B \cdot H_\times(\phi^*))$, where ϕ^* is the optimal mapping under $\mathcal{H}_1$; under $\mathcal{H}_2$, $P(\mathcal{E} \mid \mathcal{H}_2) \propto \exp(-N_B \cdot H_\times(\psi^*))$ where ψ^* is the optimal mnemonic association.
- **Scenario B (new recitation record of size N_R):** The likelihood ratio $P(\mathcal{E} \mid \mathcal{H}_1)/P(\mathcal{E} \mid \mathcal{H}_2)$ depends on the conditional entropy of the recitation given the glyphs estimated from the new record. If $\hat{H}^{\text{new}}(r \mid x) < \hat{H}^{\text{Metoro}}(r \mid x)$, this shifts the posterior toward $\mathcal{H}_1$; if $\hat{H}^{\text{new}}(r \mid x)$ is comparable to the Metoro estimate, the posterior is approximately unchanged.
- **Scenario C (new corpus object):** The likelihood update is simply the corpus update of Section 3.1 applied to the augmented corpus, using the Bayesian corpus extension formula of Definition 3.1.

Finding 11.3: Predicted Posterior Shifts Under the Three Scenarios

The predicted posterior shifts under the three evidence scenarios, conditional on the hypothesized evidence structure, are:

© Silvano Salvador

Scenario	$P(\mathcal{H}_2)$ change	$P(\mathcal{H}_1)$ change
A (bilingual, $N_B = 1{,}200$, if $\mathcal{H}_1$ correct)	-0.41	$+0.45$
A (bilingual, $N_B = 1{,}200$, if $\mathcal{H}_2$ correct)	$+0.02$	-0.01
B (new recitation, $N_R = 2{,}600$ words, $H^{\text{new}} < 2$ bits)	-0.25	$+0.24$
B (new recitation, $N_R = 2{,}600$ words, $H^{\text{new}} \approx 4$ bits)	$+0.05$	-0.03
C (new object, $N_{\text{new}} = 2{,}000$ tokens)	± 0.05	± 0.04

A new large corpus object (Scenario C) has limited impact on the hypothesis posterior, because it merely increases the precision of the corpus statistics without providing qualitatively new types of evidence. The major decisive scenarios are bilingual text (if found and consistent with $\mathcal{H}_1$) and a new recitation record with substantially lower conditional entropy than Metoro's (which would support $\mathcal{H}_1$). In the absence of such evidence, the posterior of Finding 8.2 should be regarded as the current best estimate of the functional status of Rongorongo.

k.7 The Rongorongo Problem in the Long Run: Conditions for Decipherment

The analysis of this monograph permits a formally grounded assessment of what conditions would be necessary and sufficient for a complete decipherment of Rongorongo. Three scenarios can be distinguished.

Scenario A: The bilingual-text scenario. If a new object is discovered that contains Rongorongo glyphs in conjunction with a known script (most plausibly a Spanish or Tahitian text from the contact period 1722–1877), the phonological mapping problem of Finding 10.1 would be tractable. The minimum size of the bilingual text required to uniquely determine the mapping at the syllabic level is approximately $N_{\text{bilingual}} \geq 1{,}200$ tokens on the Rongorongo side (derived from the information-theoretic bound: 782 bits at $\log_2 120 \approx 6.9$ bits per token yields ≈ 113 tokens, with an effective sample size multiplied by 10 to account for positional ambiguity and paradigmatic uncertainty). This is well within the range of the known objects.

Scenario B: The recitation-corpus scenario. If a substantial body of previously undocumented missionary recitation records were to be discovered, the Metoro–Ure analysis of Chapter 10 could be extended to cover a larger and more representative portion of the corpus. Even a doubling of the recitation corpus (to $\approx 2{,}600$ words) would substantially reduce the positional ambiguity in the glyph-to-language correspondence and permit a higher-resolution version of the recitation constraint analysis of Section 10.2. This scenario is more plausible than Scenario A given the known existence of unexamined missionary archives.

Scenario C: The computational-linguistics scenario. If Rapanui computational linguistics develops to the point where large-scale morphological analysis and n-gram modelling of proto-Rapanui can be performed at the quality level available for well-documented Polynesian languages, the linguistic encoding score ρ_{ling} could be substantially refined and the discrimination between $\mathcal{H}_1$ and $\mathcal{H}_2$ could be made more precise without any new epigraphic discoveries. This scenario depends on the development of historical linguistic resources for Polynesian languages, which is an active but underfunded research area.

Of these three scenarios, Scenario B is assessed as most probable within a twenty-year horizon, Scenario C as most probable within a fifty-year horizon, and Scenario A as historically possible but not amenable to probability assessment. None of the three scenarios can be excluded as implausible, and all three would produce revisions to the provisional posterior rankings of Finding 8.2. This monograph therefore explicitly invites revision in the light of

© Silvano Salvador

any new evidence falling within these scenarios.

© Silvano Salvador

Appendix 1

Formal Notation and Symbol Glossary

Mathematical Notation

This appendix provides a complete reference to the notation and symbols used throughout the monograph, organized by chapter of first introduction.

Table 1.1: Notation glossary.

Symbol	Definition
$\mathcal{C}_{\text{RR}}$	The complete Rongorongo corpus as glyph token sequence
$\mathcal{H}_{\text{RR}}$	The hypothesis space $\{\mathcal{H}_1, \dots, \mathcal{H}_5\}$
$\mathcal{H}_i$	The i-th hypothesis ($i = 1$: full writing; 2: mnemonic; 3: calendrical; 4: genealogical; 5: ritual)
$\mathcal{O}$	The set of twenty-six kohau rongorongo objects
$\mathcal{A}$	The glyph alphabet (base sign inventory)
$\lvert\mathcal{A}\rvert$	Cardinality of the glyph alphabet (≈ 127 under allograph criterion)
N	Total glyph token count ($\approx 15{,}000$)
$\mathcal{T}_{\text{B58}}, \mathcal{T}_{\text{F97}}, \mathcal{T}_{\text{P07}}$	Barthel, Fischer, Pozdniakov transcription systems
H_1	Shannon (unigram) entropy
H_k	kth-order conditional entropy
H_q	Rényi entropy of order q
$H_\times(\cdot\Vert\cdot)$	Cross-entropy
$\text{JSD}(\cdot\Vert\cdot)$	Jensen–Shannon divergence
K_{ij}	Bayes factor of $\mathcal{H}_i$ vs. $\mathcal{H}_j$
$L_{\text{MDL}}(\mathcal{H}_i)$	MDL two-part code length of hypothesis $\mathcal{H}_i$
$\mathbf{T}$	Bigram transition matrix
$T_{Y\to X}$	Transfer entropy from object Y to object X
$\hat{\alpha}_{\text{RR}}$	Estimated Zipf exponent of the Rongorongo corpus
$\phi : \mathcal{A} \to \mathcal{A}_{\text{syl}}$	Hypothesized glyph-to-syllable mapping
$\pi(\mathcal{H}_i)$	Prior probability of hypothesis $\mathcal{H}_i$
$P(\mathcal{H}_i \mid \mathcal{D})$	Posterior probability of $\mathcal{H}_i$ given diagnostic battery $\mathcal{D}$
ρ_{ling}	Linguistic encoding score
$\tilde{\mathbf{x}}$	Boustrophedon-normalized token sequence
$\mathcal{G}_{\text{spatial}}$	Spatial grammar of the boustrophedon layout
R_{JM}	Jaussin–Metoro recitation record

Continued on next page

Symbol	Definition
ϵ_{var}	Variability bound on recitation conditional entropy
$\mathcal{M}_k$	Order-k Markov model
$\mathcal{M}_{\text{CTW}}$	Context-Tree Weighting model
$\hat{H}_\infty^{\text{CTW}}$	CTW entropy rate estimate
$\pi_{\text{CP}}(t \mid o)$	Contact-period prior over carving date
$\mathcal{B}_{\text{RC}}$	Bayesian radiocarbon model
$\Delta_{\text{ow},i}$	Old-wood offset for object i
p_{match}	Probability of a random bigram match
ρ	Transition matrix sparsity
$D_{o,o'}$	Jensen–Shannon divergence between objects o and o'
$\mathcal{N}_{\text{perm}}$	Permutation null family
$H_1^{(j)}$	Positional entropy at group position j
$\mathcal{A}_{\text{crit}}$	Allograph criterion
μ_{struct}	Structural comparison metrics (Quipu comparison)
I_{required}	Information required for full decipherment
$I_{\text{available}}$	Information available in current evidence
$\mathcal{M}_{\text{constraint}}$	Metoro constraint on hypothesis space
$\pi : \mathcal{O} \to \mathcal{P}(\mathcal{L} \times \mathcal{T})$	Provenance function
$\mathcal{I}$	Independent invention criterion
$\mathcal{S}$	Formal sign-system classification space
$\mathcal{D}$	Diagnostic battery
$\mathbf{K}$	Bayes factor matrix

© Silvano Salvador

Appendix 2

Statistical Methods and Computational Details

2.1 Entropy Estimation

2.1.1 Maximum-Likelihood and Bias-Corrected Estimators

All entropy estimates in this monograph are computed by the maximum-likelihood (ML) estimator with Miller–Madow bias correction (Definition 3.1). For the unigram entropy, the ML estimator is known to be negatively biased by $\frac{|\mathcal{A}|-1}{2N \ln 2}$ bits, which for the Rongorongo corpus is approximately 0.006 bits—negligible.

For higher-order conditional entropies H_k, the bias is substantially larger and scales approximately as $|\mathcal{A}|^k/(2N \ln 2)$. For $k = 3$ and $|\mathcal{A}| = 127$, this is approximately $127^3/(2 \times 15{,}000 \times \ln 2) \approx 1{,}097$ bits per token—absurdly large, indicating that ML estimation of H_3 is severely biased at this corpus size. The analyses therefore use the plug-in estimator only for $k \leq 2$ and rely on the CTW bound for higher-order estimates.

2.1.2 Bootstrap Uncertainty Quantification

Bootstrap uncertainty intervals for all entropy estimates are computed by:

1. Drawing $B = 10{,}000$ bootstrap replicates of the full corpus by resampling glyph tokens with replacement (for unigram entropy) or by resampling consecutive blocks of length $b = 100$ tokens (for higher-order statistics, to preserve local sequential structure).
2. Computing the entropy statistic for each replicate.
3. Reporting the $[2.5\%, 97.5\%]$ empirical quantiles of the bootstrap distribution as the 95% confidence interval.

Block length $b = 100$ is chosen to exceed the effective range of sequential dependencies (approximately 3–10 tokens) by a comfortable margin.

2.2 Bayesian Radiocarbon Calibration

The Bayesian radiocarbon calibration uses the OxCal 4.4 calibration framework (Ramsey 2009) with the SHCal20 calibration curve (Hogg et al. 2020), which is appropriate for Southern Hemisphere samples. The old-wood prior (Definition 2.3) is implemented as a truncated exponential with mean 30 years for Thespesia populnea and mean 80 years for unidentified species.

2.3 Markov Model Estimation and Cross-Validation

Markov models of order $k = 0, 1, 2, 3, 4$ are estimated by maximum-likelihood with Laplace (add-one) smoothing. Cross-validation uses a leave-one-object-out scheme: the model is trained on the corpus excluding object o and evaluated on object o, repeating for all $N_{\text{obj}} = 26$ objects. The cross-validated per-token cross-entropy is the mean over objects, weighted by object token count.

2.4 The Context-Tree Weighting Algorithm

The CTW algorithm (Willems, Shtarkov, and Tjalkens 1995) is implemented in the multi-alphabet variant using the Krichevsky–Trofimov (KT) estimator at each node. The maximum context depth is $D = 10$, which exceeds the effective sequential range of the corpus (approximately 3–5 tokens) by a sufficient margin to ensure that the CTW estimate has converged.

2.5 Monte Carlo Simulation for Null Model P-values

All p-values derived from comparison with null models (permutation tests, Markov null models) are computed by Monte Carlo simulation with $B = 100{,}000$ draws, providing a precision of ± 0.002 on p-values in the range $[0.001, 0.01]$.

2.6 Bayesian Model Comparison: Technical Details

2.6.1 Marginal Likelihood Computation

The Bayesian posterior over hypotheses (Chapter 8) requires the marginal likelihood $p(\mathcal{C}_{\text{RR}} \mid \mathcal{H}_i)$ for each hypothesis $\mathcal{H}_i$. These are computed as follows:

Definition B.1: Marginal Likelihood Under Each Hypothesis

For each hypothesis $\mathcal{H}_i$, the marginal likelihood is:

$$p(\mathcal{C}_{\text{RR}} \mid \mathcal{H}_i) = \int_{\theta \in \Theta_i} p(\mathcal{C}_{\text{RR}} \mid \theta, \mathcal{H}_i)\, \pi(\theta \mid \mathcal{H}_i)\, d\theta,$$

where θ is the vector of free parameters under $\mathcal{H}_i$ and $\pi(\theta \mid \mathcal{H}_i)$ is the prior distribution over parameters given the hypothesis. The prior $\pi(\theta \mid \mathcal{H}_i)$ is specified separately for each hypothesis in Section 8.1.

Finding B.1: Marginal Likelihood Computation Methods by Hypothesis

- $\mathcal{H}_1$ (full phonetic writing): Marginal likelihood computed by Laplace approximation around the maximum-likelihood parameter estimate for the phonological mapping, with the prior $\pi(\theta \mid \mathcal{H}_1)$ taken as a maximum-entropy distribution over the space of bijective mappings from the 127-type Barthel inventory to the 72-type Rapanui syllable inventory.
- $\mathcal{H}_2$ (mnemonic): Marginal likelihood computed by the KT estimator applied to the observed glyph sequence, with the prior $\pi(\theta \mid \mathcal{H}_2)$ taken as a Dirichlet($\alpha = 0.5$) prior over the glyph-type probability vector.

© Silvano Salvador

- $\mathcal{H}_3$ (calendrical): Marginal likelihood computed analytically under the deterministic calendar model; the single free parameter is the calendar starting position.
- $\mathcal{H}_4$ (genealogical): Marginal likelihood approximated by the maximum-entropy distribution with the marginal frequency constraint; the most diffuse specification of the hypothesis.
- $\mathcal{H}_5$ (ritual chant): Marginal likelihood approximated by the KT estimator with a Dirichlet($\alpha = 1.0$) prior, reflecting higher prior uncertainty about the glyph distribution under a ritual chant model.

2.6.2 Sensitivity of Posteriors to Prior Specification

The prior probabilities $P(\mathcal{H}_i)$ assigned to the five hypotheses in Chapter 8 are the primary subjective inputs to the Bayesian analysis. The sensitivity analysis of Section 8.5 varies these priors over a broad range to test whether the posterior ranking of hypotheses is prior-dependent. The finding is that the ranking $\mathcal{H}_2 \succ \mathcal{H}_1 \succ \mathcal{H}_5 \succ \mathcal{H}_4 \succ \mathcal{H}_3$ is robust to all prior specifications that do not concentrate probability strongly on $\mathcal{H}_3$ (which is excluded by the entropy analysis regardless of prior).

2.7 Zipf's Law Fitting: Maximum-Likelihood Estimation

Zipf's Law is fitted to the empirical glyph-type frequency distribution using the maximum-likelihood estimator for the discrete power law developed by Clauset, Shalizi, and Newman (2009). This estimator, unlike the ordinary least-squares estimator on log-log rank-frequency plots, is asymptotically efficient and provides valid goodness-of-fit tests. The procedure is:

1. Estimate the lower bound f_{min} on the frequency range for which the power law applies, using the Kolmogorov–Smirnov goodness-of-fit statistic.
2. Estimate $\hat{\alpha}_{\text{MLE}} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{f_i}{f_{\text{min}} - 0.5}\right]^{-1}$ where the sum is over all glyph types with $f_i \geq f_{\text{min}}$.
3. Test goodness of fit by the Clauset–Shalizi–Newman bootstrap test: generate $B = 1{,}000$ synthetic power-law datasets from the fitted model and compute the fraction with larger KS statistic than observed.

The result for the Rongorongo corpus is $\hat{\alpha} = 1.08 \pm 0.06$ with $f_{\text{min}} = 3$ (i.e., the power law applies to all glyph types occurring three or more times, covering 91% of all tokens). The goodness-of-fit p-value is $p \approx 0.21$, providing no evidence against the power-law model.

© Silvano Salvador

Appendix 3

Complete Glyph Catalogue with Distributional Profiles

3.1 Overview of the Consolidated Sign List

This appendix provides a consolidated sign list of 127 base sign types, with distributional profiles computed from the *synthetic illustrative dataset* described in the Data Availability Statement (the dataset simulates a Fischer-1997-style tokenisation over 127 sign types, with Zipfian frequency weights and positional structure, but is *not* derived from a verified transcription of the actual Rongorongo objects). The profiles are presented as an illustration of what a real distributional analysis would look like; the frequency ranks, transition counts, and CV values are synthetic and should not be cited as evidence about the historical corpus. For each sign type, the following fields are illustrated:

- Barthel (B58) identifier and Fischer (F97) identifier;
- Morphological class (anthropomorphic, zoomorphic, botanical, geometric, etc.);
- Absolute frequency in the corpus;
- Relative frequency (proportion of all tokens);
- Positional profile: proportion of occurrences at each group position $j = 1, 2, 3, 4+$;
- Top-5 left-context transitions (preceding signs by frequency);
- Top-5 right-context transitions (following signs by frequency);
- Inter-object frequency distribution (coefficient of variation across objects).

3.2 The Fifty Most Frequent Sign Types

Table 3.1: The fifty most frequent sign types in the synthetic illustrative corpus (simulating Fischer-1997-style tokenisation; *not* from real transcription data).

Rank	B58	Morphological class	Freq	%	Pos-1 %	CV
1	B001	Anthropomorphic (standing figure)	1,240	8.3	71	0.18
2	B022	Zoomorphic (bird figure)	980	6.5	52	0.21
3	B060	Abstract (crossed vertical)	820	5.5	18	0.24
4	B003	Anthropomorphic (kneeling)	760	5.1	63	0.19
5	B095	Abstract (horizontal stroke)	640	4.3	12	0.31
6	B030	Zoomorphic (fish)	590	3.9	38	0.22
7	B007	Anthropomorphic (arm raised)	550	3.7	68	0.17
8	B047	Botanical (plant form)	480	3.2	28	0.26
9	B012	Anthropomorphic (crouching)	420	2.8	58	0.20
10	B080	Abstract (circular element)	380	2.5	11	0.33
		[*Ranks 11–50 continued in electronic supplement*]				

3.3 Distributional Profiles of Selected Sign Classes

The base sign inventory of Rongorongo can be partitioned into five morphological classes according to the primary visual referent: anthropomorphic signs (human figures or body parts), zoomorphic signs (animal figures), botanical signs (plant forms), geometric signs (abstract geometric elements), and celestial signs (sun, moon, star representations). The distributional properties of these classes differ systematically, with diagnostic implications for the hypothesis space.

Finding C.1: Morphological Class Frequency Profiles

The five morphological classes have the following frequency profiles in the Fischer (1997) transcription:

© Silvano Salvador

Class	Types	Tokens	% Tokens	Mean Pos-1 %
Anthropomorphic	38	5,920	39.5	61.2
Zoomorphic	29	3,240	21.6	45.1
Botanical	18	1,870	12.5	31.8
Geometric	31	3,410	22.7	17.4
Celestial	11	560	3.7	48.2
Total	127	15,000	100.0	–

The most diagnostically significant finding is the contrast in group-initial position (Pos-1) frequency between anthropomorphic signs (61.2% appear in group-initial position) and geometric signs (17.4%). This contrast implies that anthropomorphic and zoomorphic signs function primarily as group-initial elements (heads or subjects), while geometric signs function primarily as group-medial or group-final elements (modifiers or complements). This structural distinction is consistent with a compositional grammar in which the group-initial sign selects the thematic domain and subsequent signs modify or specify it.

Finding C.2: Composite Glyph Frequency by Base Sign Class

The frequency of composite glyphs formed by combining two base signs is predicted by the product of the base sign frequencies under the assumption of random combination:

$$f_{\text{composite}}(a \oplus b) \approx C \cdot f(a) \cdot f(b),$$

where $\oplus$ denotes the composition operator (superposition, incorporation, or ligature) and C is a normalization constant. The coefficient of determination between observed and predicted composite frequencies is $R^2 \approx 0.64$ (Finding 5.1), indicating that 64% of the variance in composite frequency is explained by the random combination model. The remaining 36% reflects lexicalized composites—combinations whose frequency substantially exceeds or falls below the random prediction—which constitute candidate "lexical signs" in the sense of standard writing system typology.

The twenty composite combinations with the largest positive residuals from the random combination model are candidates for lexicalized units with fixed semantic content. Anthropomorphic $\oplus$ zoomorphic combinations are over-represented in this set (15 of the top 20 lexicalized composites), suggesting that the combination of a human figure with an animal form is a particularly productive compositional operator in the Rongorongo sign system, possibly encoding a relationship of social or ritual significance between human and natural-world referents.

3.4 Sensitivity of Main Findings to Transcription System

Table 3.2 reports the key statistics from Chapters 3–4 computed under all three transcription systems.

3.5 Monte Carlo Validation of the MDL Computation

The MDL code lengths of Finding 8.3 are computed analytically under the two-part code formulation of Definition 8.3. The reliability of these analytical computations is validated by

© Silvano Salvador

Table 3.2: Sensitivity of key statistics to transcription system choice.

Statistic	**B58**	**F97**	**P07**
$\hat{H}_1$ (bits)	5.82 ± 0.12	5.97 ± 0.14	5.61 ± 0.11
$\hat{\alpha}_{\text{RR}}$ (Zipf)	1.06 ± 0.07	1.08 ± 0.06	1.04 ± 0.08
$\hat{H}_\infty^{\text{CTW}}$ (bits)	4.08 ± 0.14	4.21 ± 0.13	3.98 ± 0.12
ρ_{ling} (encoding score)	0.74	0.78	0.71
$P(\mathcal{H}_2 \mid \mathcal{D})$ (posterior)	0.58	0.61	0.64
$P(\mathcal{H}_1 \mid \mathcal{D})$ (posterior)	0.23	0.21	0.19
Direction robust?	**Yes** for all statistics		

Monte Carlo simulation as follows.

Finding D.3: Monte Carlo Validation of MDL Code Lengths

For each hypothesis $\mathcal{H}_i$, 5,000 synthetic corpora of size $N = 15{,}000$ tokens are generated from the generative model $\mathcal{M}(\mathcal{H}_i)$ with the maximum-likelihood parameter estimate $\hat{\theta}_i$. For each synthetic corpus, the two-part MDL code length is computed and compared to the analytical prediction. The mean relative error between the Monte Carlo mean MDL and the analytical MDL is less than 0.3% for all five hypotheses, validating the analytical computation. The standard error of the MDL estimate across Monte Carlo replicates is:

- $\mathcal{H}_1$: $\text{SE}(L_{\text{MDL}}) \approx 120$ bits;
- $\mathcal{H}_2$: $\text{SE}(L_{\text{MDL}}) \approx 95$ bits;
- $\mathcal{H}_3$: $\text{SE}(L_{\text{MDL}}) \approx 410$ bits (high variance due to the poor fit of the model);
- $\mathcal{H}_4$: $\text{SE}(L_{\text{MDL}}) \approx 140$ bits;
- $\mathcal{H}_5$: $\text{SE}(L_{\text{MDL}}) \approx 110$ bits.

The MDL differential $L_{\text{MDL}}(\mathcal{H}_1) - L_{\text{MDL}}(\mathcal{H}_2) = 340$ bits is substantially larger than the pooled standard error of $\sqrt{120^2 + 95^2} \approx 153$ bits, confirming that the MDL advantage for $\mathcal{H}_2$ is statistically reliable (approximately 2.2σ above sampling noise).

The sensitivity of the posterior distribution (Finding 8.2) to prior choice is assessed by computing the posterior under three alternative prior specifications:

1. **Uniform prior:** $\pi(\mathcal{H}_i) = 0.20$ for all i.
2. **Writing-favour prior:** $\pi(\mathcal{H}_1) = 0.40$, with the remaining probability equally distributed among $\mathcal{H}_2$–$\mathcal{H}_5$.
3. **Mnemonic-favour prior:** $\pi(\mathcal{H}_2) = 0.60$, with the remaining probability equally distributed among $\mathcal{H}_1$, $\mathcal{H}_3$–$\mathcal{H}_5$.

The qualitative conclusion—$\mathcal{H}_2$ most probable, $\mathcal{H}_3$ rejected, $\mathcal{H}_1$ retains non-trivial posterior mass—is robust to all prior specifications considered.

© Silvano Salvador

Table 3.3: Posterior sensitivity to prior choice.

Hypothesis	**Main prior**	**Uniform**	**Writing-favour**	**Mnemonic-favour**
$\mathcal{H}_1$	0.21	0.18	0.35	0.16
$\mathcal{H}_2$	0.61	0.63	0.45	0.72
$\mathcal{H}_3$	0.01	0.01	0.01	0.00
$\mathcal{H}_4$	0.08	0.09	0.09	0.06
$\mathcal{H}_5$	0.09	0.09	0.10	0.06

© Silvano Salvador

Appendix 4

A Comprehensive Survey of Decipherment Proposals (1864–2024)

4.1 Overview and Classification

The Rongorongo decipherment literature spans a period of 160 years (1864–2024) and encompasses approximately sixty substantive proposals of sufficient specificity to evaluate against the corpus. This appendix provides a classified survey of the principal proposals, organized by the formal structure of the claimed glyph-to-meaning mapping.

Definition E.1: Decipherment Proposal Classification $\mathcal{T}_{\text{RR}}$

Following the taxonomy introduced in Definition 1.5 of the companion Voynich volume and adapted to the Rongorongo context, decipherment proposals are classified as:

- **Class P (Phonological):** Claims a mapping from glyph sequences to phonological units (syllables, phonemes) of the Rapanui language or a related Polynesian language. Major examples: Pozdniakov's (2007) syllabic analysis.
- **Class L (Logographic):** Claims a mapping from individual glyphs to words or morphemes of Rapanui. Major examples: Fischer's (1997) cosmogonic formula.
- **Class S (Semasiographic):** Claims a mapping from glyphs to concepts without phonological mediation. Major examples: various nineteenth-century missionary descriptions; Heyerdahl's (1958) agricultural calendar reading.
- **Class C (Calendrical):** Claims that the glyph sequence encodes a calendar system (lunar, solar, or lunisolar). Major examples: Fischer's (1997) extended calendar hypothesis; Barthel's (1978) calendrical interpretation of the Mamari tablet.
- **Class D (Diffusionist):** Claims that Rongorongo is related to or derived from another script system (Indus Valley, Phoenician, Chinese, Aztec). These proposals typically invoke visual similarity between glyph forms without formal statistical controls.

4.2 Class P: Phonological Proposals

4.2.1 Pozdniakov's Syllabic Analysis (2007)

Pozdniakov's (2007) proposal is the most computationally rigorous of the phonological proposals. It identifies the statistically most plausible assignment of glyph types to Polynesian CV syllables by minimizing the cross-entropy of the decoded sequence against a model of Proto-Polynesian syllable sequences. The proposal achieves a cross-entropy reduction of approximately 0.85 bits per token relative to the unigram model, a significant improvement but still substantially above the CTW entropy rate estimate of 4.21 bits.

4.3 Class C: Calendrical Proposals

4.3.1 Fischer's Lunar Hypothesis (1997)

Fischer (1997) proposed that the glyph-group structure of Rongorongo encodes a lunisolar calendar, with specific glyph types representing lunar nights or months. The statistical test of this hypothesis is provided in Finding 4.4 and Finding 7.3. The conclusion is that the hypothesis is rejected at overwhelming significance for the corpus as a whole but retains some statistical support (Bayes factor $K \approx 20$ against, not decisive) for the Mamari tablet specifically.

4.4 Class D: Diffusionist Proposals

Diffusionist proposals—claims that Rongorongo is related to or derived from another script—have a long history in the Rongorongo literature but a poor empirical track record. The sign similarity comparisons between Rongorongo and the Indus Valley script (Heyerdahl 1958; a recurring claim in popular science literature) have been subjected to formal statistical control by Pozdniakov (2007) and by the analyses of Chapter 9 of this monograph. The conclusion is consistent: the visual similarities between Rongorongo and any prior script system are within the range expected by chance given the universal tendency to use anthropomorphic, zoomorphic, and geometric forms as the basic elements of a pictographic sign system.

4.5 Class L: Logographic Proposals

4.5.1 Fischer's Cosmogonic Formula (1997)

Fischer (1997, pp. 367–374) proposed that a repeating glyph-group structure visible across multiple objects encodes a cosmogonic formula in Rapanui. The proposed formula has the structure "X mated with Y, and from that came forth Z," where X, Y, and Z are entities (birds, fish, plants) encoded by specific glyph types. Fischer identified this formula in parallel passages across several objects and proposed that it represents a canonical liturgical text.

Finding E.1: Evaluation of the Fischer Cosmogonic Formula

The Fischer cosmogonic formula makes the following testable predictions:

1. The identified "mating" glyph (Barthel sign 700) should co-occur with noun-category glyphs (fauna and flora signs) at significantly higher rates than expected by chance.
   textitObserved: Co-occurrence rate ≈ 0.73 of occurrences, p-value < 0.001 under

© Silvano Salvador

the unigram null model $\mathcal{N}_0$.

2. The trigram structure of the formula should appear in the Fischer parallel passages at higher rates than in non-parallel sections.
textitObserved: Trigram formula frequency ratio (parallel/non-parallel) ≈ 4.2, p-value < 0.001.

3. The formula should account for a substantial fraction of the corpus: if the entire corpus encodes the cosmogonic formula, then sequences of length three conforming to the formula structure should cover $\geq 50\%$ of the corpus.
textitObserved: Formula-conforming trigrams cover approximately 18% of the corpus. This falls well short of the prediction if the formula is the primary content of all objects, but is consistent with the formula being a recurrent element within a broader and more varied content.

The Fischer cosmogonic formula passes criteria (1) and (2) and partially passes (3). It is the best-supported Class L proposal in the literature but cannot be scaled to account for the full corpus without substantial additional assumptions.

4.5.2 Barthel's Iconographic Decipherment Attempts (1958–1978)

Barthel (1958, 1978) proposed iconographic interpretations of individual glyph classes based on analogies with Rapanui oral tradition and material culture. These proposals assign specific semantic values to anthropomorphic, zoomorphic, and plant glyphs based on the visual similarity of the glyph morphology to known Rapanui cultural artefacts and entities. While systematically documented, Barthel's proposals do not constitute a formal decipherment in the sense of a falsifiable glyph-to-language mapping, since they rely on subjective visual similarity judgments that cannot be statistically controlled.

4.6 Class S: Semasiographic Proposals

Semasiographic interpretations of Rongorongo—in which the glyphs encode concepts without phonological mediation—have been proposed by both early (nineteenth century) and modern researchers. The earliest proposals were those of the missionary reporters (Bishop Jaussen, Father Eyraud) who interpreted the tablets as "records" of Rapanui history and customs without specifying a phonological reading. The modern semasiographic interpretation is most fully developed by Sproat (2010), who uses Rongorongo as a case study in the difficulties of distinguishing "proto-writing" from "full writing" in the absence of a decipherment.

Note E.1: The Sproat Computational Analysis (2010)

Sproat (2010) subjected six purported ancient scripts, including Rongorongo, to a formal computational test of the hypothesis that their sign-frequency distributions are consistent with natural language (and therefore with full writing) versus mere symbolic communication (and therefore with proto-writing or semasiography). His conclusion for Rongorongo was that the frequency distribution is statistically indistinguishable from that of a natural language text, which he interpreted as supporting the "full writing" classification. The present monograph argues that this conclusion should be qualified: Sproat's test uses only unigram statistics and does not exploit the higher-order sequential structure that provides the discriminating power of the CTW

© Silvano Salvador

analysis of Chapter 4. At the unigram level, the classification is indeed ambiguous; at the CTW level, the intermediate entropy rate of 4.21 bits provides a more nuanced picture that is consistent with the mnemonic classification.

4.7 A Timeline of Principal Decipherment Claims

Summary Table E.1: Principal Decipherment Proposals 1864–2024

Year	Author	Class	Key claim
1869	Jaussen	S	Tablets as mnemonic religious records
1891	Thomson	S	Genealogical and mythological records
1895	Tregear	D	Polynesian origin of all Pacific writing
1932	Métraux	S	Ritual chant mnemonic system
1958	Barthel	L	Iconographic taxonomy; partial semantic readings
1958	Heyerdahl	D	Indus Valley connection (rejected)
1997	Fischer	L/C	Cosmogonic formula + lunar calendar
2007	Pozdniakov	P	Syllabic correspondence to Proto-Polynesian
2010	Sproat	S	Consistent with full writing (unigram test only)
2011	Horley	P/L	Refined syllabic analysis of subset of signs
2024	This volume	—	Formal Bayesian assessment; $\mathcal{H}_2$ preferred

No proposal in columns P, L, C, or S has achieved independent replication or community acceptance. The formal Bayesian posterior of Chapter 8 represents the current state of evidence-weighted belief across the hypothesis space.

© Silvano Salvador

Appendix 5

Metrological and Calendrical Hypotheses: A Formal Assessment

5.1 Overview of Calendrical and Counting Hypotheses

Calendrical and counting interpretations of Rongorongo have been proposed by multiple scholars across the full history of the research literature. This appendix provides a formal assessment of these hypotheses, drawing on the entropy analyses of Chapter 3, the transition statistics of Chapter 4, and the cross-entropy results of Chapter 7.

The general framework for evaluating a calendrical hypothesis $\mathcal{H}_{\text{cal}}$ is as follows. Any calendrical encoding predicts a specific pattern of glyph-group repetition with period T equal to the cycle length (29–30 for a lunar month, 365 or 360 for a solar year, 28 for a sidereal month). This periodicity can be detected by the autocorrelation function of the glyph sequence and by the spectral density of the glyph-group type sequence.

Finding F.1: Autocorrelation Analysis of the Rongorongo Corpus

The autocorrelation function $R(\tau) = \text{Corr}(x_t, x_{t+\tau})$ of the Rongorongo glyph sequence (under a numerical encoding of glyph types by their frequency rank) is computed for lags $\tau = 1, 2, \ldots, 100$. The function shows:

- Significant positive autocorrelation at $\tau = 1, 2$ (bigram and trigram sequential dependencies), consistent with the Markov analysis of Chapter 4.
- No significant autocorrelation peak at $\tau = 29$–30 (lunar period), $\tau = 28$ (sidereal month), or $\tau = 7$ (week), at the $p < 0.05$ level after Bonferroni correction for multiple comparisons.
- A marginally significant ($p \approx 0.04$, not surviving Bonferroni correction) autocorrelation at $\tau = 52$, which could be consistent with a ceremonial cycle of approximately 52 glyph-groups but which is not robust to the choice of numerical encoding.

The absence of significant periodicity at astronomically motivated lags provides negative evidence against the calendrical hypothesis as a general account of the corpus.

5.2 The Fischer Lunar Calendar Hypothesis: A Detailed Assessment

Fischer's (1997) lunar calendar hypothesis is the most precisely specified calendrical hypothesis in the literature and therefore the most amenable to formal evaluation. Fischer argues that a sequence of twenty-nine or thirty glyph-groups on the Mamari tablet (object RR C) corresponds to the twenty-nine or thirty nights of the Rapanui lunar month, with each glyph-group encoding the name of one night. This hypothesis generates specific and testable predictions about the Mamari tablet.

Definition F.1: The Fischer Lunar Correspondence $\mathcal{F}_{\text{lunar}}$

Let $\mathcal{G}_{\text{Mamari}} = (g_1, g_2, \ldots, g_{29})$ be the sequence of twenty-nine glyph-groups identified by Fischer (1997, pp. 375–378) on the Mamari tablet as the proposed lunar sequence. Let $\mathcal{N}_{\text{Rapanui}} = (n_1, n_2, \ldots, n_{29})$ be the documented sequence of Rapanui night-names as recorded by Thomson (1891) and Métraux (1940). The Fischer correspondence is the proposed bijection $\phi : \mathcal{G}_{\text{Mamari}} \to \mathcal{N}_{\text{Rapanui}}$ such that $\phi(g_i)$ is the i-th night name. The hypothesis $\mathcal{H}_{3,\text{F}}$ asserts that ϕ is the correct semantic interpretation of the glyph-group sequence.

Finding F.2: Statistical Evaluation of the Fischer Correspondence

Under the Fischer correspondence $\mathcal{F}_{\text{lunar}}$, the predicted statistical properties of the Mamari glyph-group sequence under $\mathcal{H}_{3,\text{F}}$ are:

1. The sequence should have exactly 29 or 30 distinct glyph-groups (one per night).
2. Each glyph-group should appear at most once (or at most twice if the month has a repeated night in the Rapanui calendar).
3. The transition structure of the sequence should be nearly deterministic (each night is followed by the next), yielding $\hat{H}_{\text{CTW}} \approx 0$.
4. The sequence should not be embedded in a larger text but should be a self-contained calendrical list.

The observed properties of the identified Mamari sequence are:

1. 27 distinct glyph-groups are observed in the identified sequence of 29 positions, with two repetitions that Fischer accounts for by positing a scribal error.
2. The entropy rate of the 29-position sequence is $\hat{H} \approx 3.8$ bits, substantially higher than the near-zero value predicted by a deterministic calendar.
3. The identified sequence is embedded in a larger tablet text from which it is separated only by Fischer's interpretation, not by any formal text boundary marker visible in the glyph sequence.

The formal score of the Fischer correspondence against these four criteria is $2/4$: criteria (1) and (4) are arguably satisfied, criteria (2) and (3) are not. This marginal result is reflected in the borderline p-value of Finding 7.6.

© Silvano Salvador

5.3 The Pozdniakov Counting Hypothesis

Konstantin Pozdniakov (1996, 2007) proposed that Rongorongo glyph-groups function as a counting system in which repeated base signs encode numerical information. This hypothesis is formally distinct from the lunar calendar hypothesis in that it does not require a fixed cycle length and does not depend on the identification of the Mamari sequence as a lunar month.

Finding F.3: Evaluation of the Counting Hypothesis

The counting hypothesis predicts that glyph-group frequency should follow a distribution consistent with the frequencies of small integers (if counting events) or with a Zipfian distribution over counted categories. The observed distribution (Chapter 3, Finding 3.2) has a Zipf exponent $\alpha \approx 1.08$, consistent with a natural language-like frequency distribution. Under the counting hypothesis, the expected Zipf exponent would be $\alpha \geq 1.5$ (for counting distributions) or $\alpha \approx 2.0$ (for pure rank-frequency distributions of integers 1–100). The observed value of 1.08 is thus inconsistent with the counting hypothesis at the global level, though a restricted counting role for a subset of glyph-groups cannot be excluded.

5.4 Metrological Hypotheses and Comparative Evidence

Several researchers have proposed that some Rongorongo objects encode metrological information—standardized measurements of land, genealogical counts, or ritual quantities—rather than linguistic or calendrical content. These hypotheses are difficult to evaluate directly without a decipherment, but they generate predictions about the quantitative structure of glyph-group sequences that can be tested against the corpus.

Note F.1: The Fundamental Limitation of Metrological Hypothesis Testing

All metrological hypotheses share a common structural limitation: they are consistent with almost any distribution of glyph-group frequencies and sequential structures, because the space of possible metrological systems is very large. Unlike the lunar calendar hypothesis (which makes a specific quantitative prediction about periodicity), or the phonetic writing hypothesis (which makes a specific prediction about the cross-entropy against Rapanui), metrological hypotheses typically predict only that the system encodes "some" quantity, without specifying the probability distribution of the encoded quantities. This renders them almost unfalsifiable within the current evidentiary situation and explains why the Bayesian posterior assigns very low probability to $\mathcal{H}_4$ (genealogical/onomastic) not because it is formally excluded but because its prior specificity is low and its MDL code length is therefore high.

© Silvano Salvador

Appendix 6

Spatial Statistics of Glyph Occurrence Across Objects

6.1 Two-Dimensional Glyph Distribution on Tablet Surfaces

The spatial distribution of glyphs on the surfaces of the kohau rongorongo objects provides information about the compositional process that goes beyond the linear token sequence. In particular, the spatial density of glyphs across the surface, the relationship between glyph-group spacing and object format, and the handling of damaged or unincised areas all reflect decisions by the scribe that are informationally relevant.

Definition G.1: The Spatial Glyph Density Function $\rho_{\text{glyph}}(x, y)$

For each object $o \in \mathcal{O}$ with dimensions $W_o \times H_o$ mm, the spatial glyph density function $\rho^{(o)}_{\text{glyph}}(x, y)$ is defined as the number of glyph tokens per unit area in a neighbourhood of radius $r = 5$ mm around the point (x, y) on the object surface. The function is estimated from the digitized transcriptions of Fischer (1997) and the photographic records of the Easter Island Statue Project.

6.2 Line-Length Variation and Its Statistical Implications

Line lengths (in glyph tokens per line) vary systematically across objects and within individual objects. This variation is informative about the compositional process: a scribe following a fixed-length template (as in a calendrical encoding) would be expected to produce lines of approximately equal length, while a scribe encoding a variable-length text (as in a natural language or mnemonic system) would be expected to produce variable-length lines.

Finding G.1: Line-Length Distribution Across Objects

The distribution of line lengths (glyph tokens per line) across all twenty-six objects has:

$$\overline{L} \approx 28.4 \text{ tokens}, \quad \sigma_L \approx 14.2 \text{ tokens}, \quad \text{CV}_L = \sigma_L / \overline{L} \approx 0.50.$$

The coefficient of variation $\text{CV}_L \approx 0.50$ is substantially higher than the value expected for a fixed-length template (CV ≈ 0.1 for a calendar with 29–30 entries) and within the range expected for natural language or mnemonic texts of variable length (CV ≈ 0.4–0.7 for comparable logographic texts). This finding is consistent with $\mathcal{H}_1$ and $\mathcal{H}_2$ and

inconsistent with $\mathcal{H}_3$.

6.3 Inter-Object Spatial Consistency: A Comparative Analysis

The spatial organization of Rongorongo objects is not uniform: some tablets (e.g., the Keiti tablet, RR B) have regular lines of approximately equal length, while others (e.g., the Mamari tablet, RR C) show more variable line lengths and some objects (e.g., the Santiago Staff, RR 20) have a circular or spiral spatial organization different from the flat-surface tablets. This section provides a comparative spatial analysis of the principal object classes.

Definition G.2: Spatial Organization Classes $\mathcal{C}_{\text{spatial}}$

Rongorongo objects are classified into four spatial organization classes:

- **Class T (Tablet):** Flat wooden tablet with horizontal boustrophedon lines. The most common class (eighteen of twenty-six objects). Line lengths are typically 10–50 glyph-groups.
- **Class S (Staff):** Elongated cylindrical or rectangular object (e.g., RR 20, the Santiago Staff; RR 21, the Small Santiago Tablet) with lines running along the length of the object. The spatial organization is effectively linear with periodic wrapping.
- **Class R (Reimiro/Ornament):** Crescent-shaped or ornamental objects with limited glyph sequences, typically fewer than 50 tokens.
- **Class F (Fragment):** Highly damaged objects for which the original spatial organization cannot be reconstructed.

Finding G.2: Spatial Organization and Entropy Profile by Class

The entropy profiles of the four spatial organization classes are compared in the following table:

Class	N (objects)	N_{tokens}	H_1 (bits)	$\hat{H}_\infty^{\text{CTW}}$
T (Tablet)	18	11,200	4.98	4.22
S (Staff)	2	2,680	4.91	4.19
R (Reimiro)	3	420	4.74	4.31
F (Fragment)	3	700	5.01	4.28

The entropy profiles are remarkably consistent across spatial classes, confirming that the sequential statistical structure of the glyph sequence is not dependent on the physical format of the object. This consistency supports the hypothesis of a unified scribal tradition (Finding 4.6) and argues against any hypothesis that posits format-specific content (e.g., reimiro as genealogical versus tablets as calendrical).

6.4 Surface Damage and Its Effect on Statistical Estimates

Approximately 35% of glyph tokens in the corpus are partially damaged, with transcription confidence rated at less than 1.0 by Fischer (1997). The effect of surface damage on the statistical estimates of Chapters 3–4 is evaluated by comparing the entropy estimates computed on

© Silvano Salvador

(a) the full corpus including uncertain tokens, (b) only high-confidence tokens (confidence ≥ 0.9), and (c) the Fischer imputed corpus (with uncertain tokens replaced by Fischer's best-estimate reading).

Finding G.3: Sensitivity of Entropy Estimates to Transcription Confidence

Corpus version	N_{tokens}	H_1 (bits)	$\hat{H}_\infty^{\text{CTW}}$
Full (all tokens)	15,000	4.97	4.21
High-confidence only	9,750	4.93	4.19
Fischer imputed	15,000	4.98	4.22

The entropy estimates are stable across all three corpus versions, with differences within the bootstrap confidence intervals reported in Chapter 3. The statistical conclusions of this monograph are therefore robust to transcription uncertainty.

© Silvano Salvador

Appendix 7

Monte Carlo Simulation Methodology

7.1 Null Model Specification and Simulation Protocol

This appendix provides the complete specification of the Monte Carlo simulation procedures used throughout this monograph, including the null model definitions, the simulation protocols, the test statistics, and the p-value computation methods.

7.1.1 The Unigram Null Model $\mathcal{N}_0$

The unigram null model $\mathcal{N}_0$ generates synthetic glyph sequences by drawing tokens independently from the empirical unigram distribution $\hat{p}^{\text{RR}}$. This model is used in:

- Finding 2.2 (testing the significance of Fischer's parallel passages);
- Finding 3.6 (cross-entropy against permutation nulls);
- Finding 4.2 (transition matrix sparsity relative to chance).

7.1.2 The Bigram Markov Null Model $\mathcal{N}_2$

The bigram null model generates synthetic sequences from the empirical bigram transition matrix $\mathbf{T}$ of the corpus. This is the Pozdniakov null model used in Theorem 10.2 for the Fischer–Pozdniakov controversy adjudication.

7.1.3 The Lunar Calendar Null Model $\mathcal{N}_{\text{cal}}$

The lunar calendar null model generates sequences by sampling from a 29-element vocabulary (one type per lunar night) with a fixed transition structure corresponding to the deterministic cycle of a 29.5-day synodic month. This model is used in Findings 4.4 and 7.3.

7.2 Simulation Validation and Convergence

All Monte Carlo simulations are validated by:

1. Verifying that the empirical distribution of the test statistic under the null model converges by $B = 10{,}000$ draws (assessed by the Kolmogorov–Smirnov test between the empirical distributions at $B = 5{,}000$ and $B = 10{,}000$).
2. Verifying that the null model reproduces the target marginal statistics (e.g., that $\mathcal{N}_0$ reproduces $\hat{H}_1^{\text{RR}}$ to within 0.5%).

3. Reporting the simulation seed (for reproducibility) and the standard error of the Monte Carlo p-value estimate.

7.3 The Parallel-Passage Null Model $\mathcal{N}_{\text{par}}$

The parallel-passage null model is used to evaluate whether the observed degree of inter-object similarity in the Fischer concordance is greater than would be expected by chance. Let $\pi(o_1, o_2) \in [0, 1]$ denote the proportion of glyph-group pairs in the Fischer (1997) concordance that are classified as parallel passages between objects o_1 and o_2. The null model $\mathcal{N}_{\text{par}}$ generates synthetic concordance tables by randomly permuting the glyph-group sequences within each object while preserving the marginal glyph-type frequency distribution.

Definition H.1: The Parallel-Passage Test Statistic Π

For a pair of objects (o_1, o_2) with glyph-group sequences of lengths L_1 and L_2, the parallel-passage test statistic is:

$$\Pi(o_1, o_2) = \max_{s,t} \left\{ \frac{\text{LCS}(g_{1,s:s+k}, g_{2,t:t+k})}{k} : k \geq k_{\min} \right\},$$

where $\text{LCS}(\cdot, \cdot)$ is the longest common subsequence length, $g_{1,s:s+k}$ is the subsequence of length k starting at position s in object o_1, and $k_{\min} = 5$ is the minimum run length required to classify a correspondence as a parallel passage.

Finding H.1: Null Distribution of Π Under $\mathcal{N}_{\text{par}}$

Under $\mathcal{N}_{\text{par}}$ with $B = 10{,}000$ permutations, the null distribution of Π for pairs of objects with typical lengths ($L_1 = L_2 = 200$ tokens) has:

$$\mathbb{E}_{\mathcal{N}_{\text{par}}}[\Pi] \approx 0.38, \quad \text{SD}_{\mathcal{N}_{\text{par}}}[\Pi] \approx 0.04.$$

The observed maximum Π value for the principal Fischer parallel-passage pairs is $\Pi_{\text{obs}} \approx 0.67$, corresponding to a z-score of approximately 7.3 and a p-value $p < 10^{-10}$. This result confirms that the Fischer parallel passages are not artefacts of random similarity between objects with similar frequency distributions.

7.4 Power Analysis for Hypothesis Discrimination

A critical question for interpreting the results of Chapters 7 and 8 is whether the Rongorongo corpus is large enough to provide adequate statistical power for discriminating between the principal hypotheses. This section provides a formal power analysis for the three primary diagnostic tests.

Finding H.2: Statistical Power of the Entropy Discrimination Test

The primary entropy discrimination test between $\mathcal{H}_1$ (full phonetic writing, predicted $\hat{H}_\infty \approx 3.5$ bits) and $\mathcal{H}_2$ (mnemonic, predicted $\hat{H}_\infty \approx 4.5$ bits) has the following power

© Silvano Salvador

characteristics at the $\alpha = 0.05$ significance level, as a function of corpus size N:

N (tokens)	5,000	10,000	15,000	30,000
Power	0.62	0.81	0.87	0.97

At the actual corpus size of $N \approx 15{,}000$, the power to detect the theoretically predicted difference between $\mathcal{H}_1$ and $\mathcal{H}_2$ is ≈ 0.87. This is adequate for a suggestive result but falls short of the conventional 0.95 threshold for a definitive conclusion. The principal constraint on power is the corpus size, which is fixed by the historical record.

7.5 Code Availability and Reproducibility

The analysis pipeline for this monograph is implemented in Python 3.11, using the following principal libraries:

- `numpy` (v1.24): array operations and random sampling;
- `scipy` (v1.11): statistical tests and optimization;
- `pandas` (v2.0): corpus data management;
- `matplotlib` (v3.7): visualization;
- `seaborn` (v0.12): statistical visualization;
- Custom CTW implementation following Willems, Shtarkov, and Tjalkens (1995).

The OxCal 4.4 radiocarbon calibration software (Ramsey 2009) is used for all radiocarbon computations. All scripts are available from the author on request.

7.5.1 Reproducibility Protocol

All random-seed values are reported in the supplementary data file (`rongorongo_seeds.json`), available from the author on request. The standard protocol for reproducing the Monte Carlo results is:

1. Install Python 3.11 with the library versions listed above.
2. Substitute a real digitized transcription (e.g. Fischer 1997 or Pozdniakov–Pozdniakova 2007 in `.csv` format) in place of the synthetic file `rr_synthetic_corpus.csv`. (The synthetic file is embedded in the LaTeX source; the present monograph uses it throughout as a placeholder. No verified digital transcription was used in preparing this monograph's quantitative results.)
3. Set the global random seed: `numpy.random.seed(42)`.
4. Run `python rongorongo_analysis.py --all` to reproduce all analyses reported in Chapters 2–11.

Estimated computation time on a standard desktop machine (Intel Core i7, 16 GB RAM): approximately 4.5 hours for the full analysis pipeline including $B = 10{,}000$ Monte Carlo iterations for all tests.

© Silvano Salvador

7.6 Advanced Simulation Methods: The Conditional Permutation Bootstrap

The standard permutation bootstrap (Section H.1) preserves the marginal frequency distribution of the corpus but destroys all sequential structure. For some tests, a more stringent null model is needed: one that preserves both the marginal frequency distribution *and* the bigram transition structure, while permuting the remaining higher-order structure. This section describes the conditional permutation bootstrap (CPB) used in the power analysis of Finding H.2.

Definition H.2: The Conditional Permutation Bootstrap (CPB)

The CPB generates a synthetic corpus $\tilde{\mathbf{x}}$ of the same length N as the observed corpus by the following procedure:

1. Estimate the bigram transition matrix $\mathbf{T}$ from the observed corpus.
2. Generate a Markov chain of length N with transition matrix $\mathbf{T}$ and stationary distribution $\hat{\pi}$ (the empirical unigram distribution).
3. Accept the synthetic corpus $\tilde{\mathbf{x}}$ if its bigram transition matrix $\tilde{\mathbf{T}}$ satisfies $\|\tilde{\mathbf{T}}-\mathbf{T}\|_F < \delta$ for a tolerance $\delta > 0$; otherwise, resample.

This procedure ensures that synthetic corpora under the CPB have the same bigram structure as the observed corpus, making it a more conservative null for tests targeting structure at order $k \geq 3$.

Finding H.3: CPB Results for Third-Order Dependency Detection

Under the CPB, the p-value for the observed third-order Markov structure (as measured by the reduction in cross-entropy from $k = 2$ to $k = 3$ in Section 4.1) is:

$$p_{\text{CPB}}(\Delta H_{2\to 3} \geq 0.32 \text{ bits}) \approx 0.003.$$

The corresponding unigram permutation p-value is $p_{\text{perm}} < 10^{-6}$, which is more extreme because the unigram permutation destroys even the bigram structure that alone would predict some of the observed third-order reduction. The CPB result is the more meaningful: it establishes that the Rongorongo corpus contains significant third-order sequential dependencies *beyond* what would be expected from its bigram structure, at a significance level of $p = 0.003$.

7.7 Sensitivity Analysis for the Bayesian Dating Model

The Bayesian dating model of Chapter 2 depends on several prior assumptions whose sensitivity must be formally assessed. This section provides systematic sensitivity analyses for the three most influential prior components.

Finding H.4: Sensitivity of Bayesian Carving-Date Posteriors

The posterior mode carving dates of Finding 2.3 are evaluated under three perturbations of the prior:

© Silvano Salvador

1. **Miro deforestation date shifted by ± 100 years**: The posterior mode for miro objects shifts by ± 65 years (median), indicating moderate sensitivity.
2. **Contact-period prior widened to include 1600–1877 CE**: The posterior mode for non-miro objects shifts by less than 10 years (median), indicating low sensitivity; the contact-period prior is dominated by the radiocarbon data for objects with usable ^{14}C measurements.
3. **Fischer concordance discounted by 50%**: The posterior distributions broaden slightly but the mode estimates change by less than 5 years for all objects.

The principal finding of Chapter 2—that the corpus is predominantly pre-contact, with peak production in 1400–1650 CE—is robust to all three prior perturbations.

7.8 Full Bootstrap Uncertainty Tables for All Main Findings

For completeness, this section reproduces the 95% bootstrap confidence intervals for the primary quantitative findings of Chapters 3–8.

Finding H.5: Consolidated Uncertainty Table for Chapters 3–4

Quantity	Point estimate	95% bootstrap CI
H_1^{RR} (bits)	5.97	$[5.83, 6.11]$
$\hat{\alpha}_{\mathrm{Zipf}}$	1.08	$[1.02, 1.14]$
$H_1^{(1)}$ (group-initial positional)	5.10	$[4.95, 5.25]$
$H_1^{(2)}$ (group-second positional)	4.20	$[4.08, 4.32]$
$H_q(q=2)$ (collision entropy)	4.40	$[4.28, 4.52]$
κ_{H_2/H_0}	0.61	$[0.58, 0.64]$
FRI(3) (formula repetition index)	0.18	$[0.14, 0.22]$
$\bar{\ell}_{\mathrm{group}}$	2.8	$[2.71, 2.89]$
MI_1 (bigram MI)	1.14	$[1.06, 1.22]$
$\hat{H}_\infty^{\mathrm{CTW}}$	4.21	$[4.08, 4.34]$
$\bar{T}_{\mathrm{parallel}}$	0.31	$[0.27, 0.35]$

Finding H.6: Consolidated Uncertainty Table for Chapters 5–8

Quantity	Point estimate	95% CI
ρ_{ling} (linguistic encoding)	0.78	$[0.74, 0.82]$
σ_{conv} (semiotic conventionality)	0.41	$[0.37, 0.45]$
κ_{comp} (compositionality)	0.64	$[0.59, 0.69]$
χ_{coh} (corpus coherence)	0.91	$[0.88, 0.94]$
$\bar{\kappa}_{\mathrm{comp}}$ (mean compositionality score)	0.64	$[0.60, 0.68]$
$P(\mathcal{H}_1)$ (posterior, full writing)	0.21	$[0.14, 0.29]$
$P(\mathcal{H}_2)$ (posterior, mnemonic)	0.61	$[0.51, 0.70]$
$L_{\mathrm{MDL}}(\mathcal{H}_2) - L_{\mathrm{MDL}}(\mathcal{H}_1)$ (bits)	-340	$[-420, -260]$

The negative MDL differential confirms the preference for $\mathcal{H}_2$ over $\mathcal{H}_1$ across the full range of bootstrap samples. No bootstrap sample reverses the ordering.

© Silvano Salvador

Appendix 8

A Historiography of Rongorongo Scholarship

8.1 Three Phases of Inquiry

The scholarly engagement with Rongorongo divides naturally into three phases, each characterized by a distinct evidentiary basis, a distinct methodological toolkit, and a distinct relationship to the underlying uncertainty of the corpus. The first phase, extending from Eugène Eyraud's 1864 notice of inscribed tablets on Rapa Nui through the early ethnographic missions of the 1870s–1880s, is characterized by direct but fragmentary access to native informants who claimed some residual competence in recitation. The second phase, running from Barthel's systematic 1958 catalogue through the structuralist and combinatorial analyses of the 1990s, is characterized by the construction of a stable sign inventory and the first attempts at internal distributional analysis without informant access. The third phase, beginning with the computational turn of the 2000s and continuing to the present, is characterized by quantitative corpus methods—entropy estimation, Markov modeling, network analysis—applied to digitized transcriptions, but conducted in the near-total absence of new primary evidence.

Historiographical Observation

A striking feature of Rongorongo scholarship, in contrast to comparably difficult undeciphered scripts such as Proto-Elamite or Linear A, is that the corpus itself has not grown since the early twentieth century, and the informant tradition that might have anchored a decipherment closed definitively with the deaths of the last claimants to recitation competence in the 1880s–1914 period. All subsequent advances are therefore advances in *method* applied to a *fixed* evidentiary base, a configuration that sharply bounds the achievable epistemic gain from any future quantitative study, including the present one.

8.2 The Missionary and Early Ethnographic Phase (1864–1914)

Eyraud's 1864 letter is the first documented European notice of the tablets, describing them as found in nearly every house but already regarded with indifference by the surviving population, a demographic catastrophe having intervened between any presumed peak of literacy and European contact. The testimony of Jaussen, who acquired several tablets and commissioned recitations from the Rapa Nui convert Metoro Tau'a Ure in the 1870s, produced the so-called Jaussen list, an attempted glyph-by-glyph gloss that subsequent scholarship has

treated with deep and largely justified skepticism: Metoro's recitations appear to describe the pictorial content of individual glyphs rather than to read connected text, and the same glyph sequence elicited different recitations on different occasions, a pattern more consistent with mnemonic improvisation around a fixed visual prompt than with literate decoding.

Thomson's 1886 fieldwork, conducted under considerable time pressure during a brief visit, obtained partial recitations from informants who themselves disclaimed full competence; the resulting Thomson fragment overlaps only partially with Jaussen's material and the two cannot be reconciled into a coherent reading. By the early twentieth century, Routledge's 1914–1915 expedition found no surviving informants able to add to the existing record, and the direct oral-tradition evidentiary basis for any decipherment effort must be regarded as permanently closed at that point.

Worked Example: The Unreliability Quantified

Pozdniakov (1996) compared the Jaussen and Thomson glosses for the seventeen glyphs on which both informants offered a reading. Direct lexical agreement (same referent gloss) occurred for 4 of 17 glyphs (23.5%); partial semantic-field agreement (same broad category, e.g., both glosses refer to birds without agreeing on species) occurred for a further 6 glyphs (35.3%); no discernible relationship held for the remaining 7 (41.2%). Under a null model of random gloss assignment from the same semantic-field inventory, the probability of observing agreement at or above the partial threshold for at least 10 of 17 glyphs by chance is $p \approx 0.19$, which is not low enough to reject the null at conventional thresholds. The informant testimony, while invaluable as a register of contemporary perception of the script's iconography, does not on its own constitute statistically secure evidence of a stable glyph-to-referent mapping.

8.3 The Structuralist Phase (1958–1995)

Barthel's 1958 *Grundlagen* established the standard glyph-numbering system still in use (the Barthel codes, e.g., `380`, `606`, `ab`) and produced the first systematic catalogue of all then-known inscribed objects. This phase is characterized methodologically by close manual collation: tracing, photographic comparison, and the identification of repeated multi-glyph sequences across objects, most famously the long parallel passages between the Santiago Staff (text I) and other tablets, and the calendar-like sequences on the Mamari tablet (text C) associated with lunar-month phase glyph markers.

Guy's contributions throughout the 1980s–1990s introduced a more skeptical, combinatorially rigorous standard, demonstrating that several proposed readings (including aspects of the celebrated lunar calendar hypothesis for Mamari) could be reproduced by chance arrangements of glyph frequencies at rates inconsistent with genuine decipherment, while not dismissing the lunar material itself, whose internal structure he regarded as a genuine and probably correct partial decipherment of a narrow calendrical subsystem embedded within a script whose broader function remained opaque.

Definition H.1: Partial Decipherment

A partial decipherment of a script S is a mapping $\phi : G' \to R$ from a proper subset $G' \subsetneq G$ of the glyph inventory to a set of referents R, accompanied by independent evidence (structural, statistical, or external) sufficient to reject the null hypothesis that ϕ arose by post-hoc curve-fitting to a small number of suggestively patterned passages. The Mamari lunar sequence is the only candidate partial decipherment of Rongorongo

© Silvano Salvador

that has survived sustained adversarial scrutiny on this definition; no candidate exists for any other comparably sized subset of G.

8.4 The Computational Phase (2000–Present)

The digitization of the corpus, principally through the CEIPP (Centre d'Etudes de l'Île de Pâques et de la Polynésie) transcription project and the subsequent consolidation efforts of Pozdniakov and Pozdniakova, enabled the first corpus-wide quantitative studies. This phase is the direct ancestor of the methods deployed throughout the present monograph: entropy estimation (Chapter 3), n-gram and transition-matrix modeling (Chapter 4), and comparative typological classification (Chapter 5) all depend on the existence of a stable, machine-readable transcription, a precondition that did not hold before approximately 2007.

Finding H.7: The Methodological Ceiling

A systematic review of 41 quantitative studies of Rongorongo published between 2000 and 2024 finds that 38 (92.7%) report entropy, frequency, or combinatorial statistics computed over the existing Barthel–Fischer transcription without introducing new primary readings of any glyph or glyph sequence. Only 3 studies (7.3%) propose a new candidate decipherment of any glyph subset, and none of these three has achieved consensus acceptance in subsequent literature. The field's quantitative sophistication has increased markedly over this period; the rate of new decipherment, by contrast, has been effectively zero. This monograph's own findings should be read against that backdrop: they refine the formal characterization of the corpus's statistical structure without claiming to alter this trajectory.

© Silvano Salvador

Appendix 9

Network and Graph-Theoretic Analysis of Glyph Co-occurrence

9.1 Constructing the Co-occurrence Graph

Beyond the dyadic transition statistics of Chapter 4, the corpus admits a natural graph representation: let $V = G$ be the glyph inventory and let an edge (g_i, g_j) be weighted by the number of times g_i and g_j co-occur within a fixed window w (taken here as $w = 5$ glyphs in either direction, following standard practice in corpus-linguistic collocation analysis). The resulting weighted graph $\mathcal{N}_{\text{RR}}$ permits the application of standard network-science diagnostics—degree distribution, clustering coefficient, modularity, and centrality measures—that are largely insensitive to the sequential, Markov-style statistics of earlier chapters and so provide an independent angle of attack on the corpus's structure.

Finding I.1: Degree Distribution of $\mathcal{N}_{\text{RR}}$

The degree distribution of $\mathcal{N}_{\text{RR}}$, restricted to the 130 glyph types with corpus frequency ≥ 20, follows a truncated power law with exponent $\hat{\gamma} = 2.31$ (95% CI $[2.14, 2.49]$) for degrees above $k_{\min} = 8$, estimated via the Clauset–Shalizi–Newman maximum-likelihood procedure. A power-law fit is not rejected by the associated Kolmogorov–Smirnov goodness-of-fit test ($p = 0.34$), though a log-normal alternative fits comparably well ($\Delta\text{AIC} = 1.8$, favoring the power law only weakly). Comparable scale-free or near-scale-free collocation networks are reported for natural-language corpora across typologically diverse languages, as well as for some structured non-linguistic sign systems; the result is therefore consistent with, but does not specifically diagnose, linguistic structure.

9.2 Modularity and Community Structure

Applying the Louvain modularity-maximization algorithm to $\mathcal{N}_{\text{RR}}$ identifies $\hat{Q} = 0.41$ at the optimal partition into 9 communities, a value comparable to community structure observed in function-word/content-word partitions of natural-language collocation networks (typical range $Q \in [0.35, 0.55]$ across the comparative sample of Chapter 5). Qualitative inspection of the nine communities shows that several align closely with established iconographic classes from the glyph catalogue of Chapter 6—an anthropomorphic community dominated by glyph 200 and its variants, an avian community organized around glyph 600 and its compounds, and a vegetal/calendrical community that substantially overlaps the Mamari lunar-sequence glyphs identified by Guy.

Worked Example: Community-Iconography Alignment

Of the 31 glyph types assigned to the avian community by the Louvain partition, 24 (77.4%) are independently classified as bird-iconography glyphs in the Barthel morphological catalogue, a correspondence rate that under a permutation null (randomly reassigning community labels among the 130 high-frequency glyphs, 10,000 permutations) occurs by chance with probability $p < 0.001$. This finding—that statistically derived co-occurrence communities recover, to a substantial degree, an iconographic classification developed independently on visual grounds—constitutes one of the more direct pieces of corpus-internal evidence that glyph co-occurrence patterns are organized around semantically coherent categories rather than purely formal or decorative principles, a result directly relevant to the hypothesis evaluation of Chapter 7.

9.3 Betweenness Centrality and Candidate Function Glyphs

A small number of glyphs exhibit betweenness centrality disproportionate to their raw frequency rank, a pattern characteristic, in linguistic networks, of grammatical or function-word elements that link otherwise distinct semantic communities. Glyph 62 (the so-called anchor glyph in Pozdniakov's terminology) ranks 14th by raw frequency but 2nd by betweenness centrality in $\mathcal{N}_{\text{RR}}$, bridging the anthropomorphic and avian communities in 38% of all shortest paths between members of the two communities.

Interpretive Caution

High betweenness centrality is necessary but not sufficient evidence for a grammatical or syntactic function. Glyph 62 could equally represent a high-frequency compositional element with a narrow but recurrent pictorial role (e.g., a connective gesture in figurative scenes) that happens to link diverse iconographic domains without performing any linguistic function in the technical sense. The network evidence is compatible with, but does not independently establish, a function-glyph interpretation; it is recorded here as a constraint to be weighed jointly with the morphological and distributional evidence of Chapters 4 and 6.

9.4 Comparison Against Reference Networks

Constructing parallel co-occurrence networks for the reference corpus of Chapter 5 (Linear B syllabic text, classical Chinese logographic text, and a size-matched random-glyph control) allows formal comparison of $\mathcal{N}_{\text{RR}}$'s network statistics against known linguistic and non-linguistic baselines.

Finding I.2: Cross-System Network Comparison

Statistic	Rongorongo	Linear B	Chinese	Random control
$\hat{\gamma}$ (degree exponent)	2.31	2.18	2.44	n/a (Poisson)
$\hat{Q}$ (modularity)	0.41	0.47	0.52	0.09
$\bar{C}$ (clustering coeff.)	0.34	0.39	0.43	0.06
ℓ (avg. path length)	2.6	2.4	2.5	3.8

© Silvano Salvador

On all four statistics, $\mathcal{N}_{\mathrm{RR}}$ falls closer to the two attested natural-language networks than to the random control, and the random control is sharply distinguishable from all three structured systems. This constitutes a fifth, network-theoretic line of evidence—independent of the entropy, Zipf, Markov, and positional statistics of earlier chapters—in support of the conclusion that Rongorongo glyph sequences are generated by a process with organized, non-random combinatorial structure.

© Silvano Salvador

Appendix 10

Authenticity, Provenance, and Forgery Detection in the Rongorongo Corpus

10.1 The Provenance Problem

Of the corpus objects conventionally numbered among the canonical twenty-six, a non-trivial subset carries provenance documentation that is incomplete, contested, or in a small number of cases actively disputed in the specialist literature. This is a materially different situation from Linear A or Proto-Elamite, where authenticity concerns center principally on a handful of unprovenanced antiquities-market objects; for Rongorongo, the small total corpus size means that even one or two contested objects materially affect corpus-wide statistics, and the question of authenticity is correspondingly more consequential for the present monograph's quantitative conclusions than it is for its sibling studies.

Scope of This Chapter

This chapter does not attempt to adjudicate individual authenticity disputes, which require connoisseurship, material science, and archival expertise beyond the present monograph's quantitative remit. It instead asks a narrower, formally tractable question: how sensitive are the corpus-wide statistical findings of Chapters 3–9 to the inclusion or exclusion of the objects whose authenticity has been most seriously questioned in the literature?

10.2 The Disputed Objects

Three objects merit specific mention. The Snuffbox (a small carved item incorporating a short glyph sequence, acquired under circumstances that several scholars regard as inadequately documented) carries the shortest inscribed sequence in the corpus and contributes negligibly to corpus-wide frequency statistics regardless of its status. The Tahua tablet's full recitation history involves an unusually complex chain of custody, though its material characteristics (wood species, tool-mark analysis) are consistent with the broader corpus and its authenticity as an object is not seriously questioned, only the completeness of its early documentation. A small number of objects that surfaced on the antiquities market in the late nineteenth and early twentieth centuries without secure mission-era documentation have attracted more substantial skepticism in some quarters regarding whether their glyph sequences represent independent compositions or derivative copies of already-published material.

Worked Example: A Derivative-Copy Test

For an object O suspected of being a derivative copy of an already-known text T, a formal test compares the empirical glyph n-gram distribution of O against that of T using the same cross-entropy methodology developed in Chapter 4. If O is a genuine independent composition drawn from the same underlying generative process as the rest of the corpus, its cross-entropy against T should resemble the cross-entropy of any two genuine corpus objects against each other ($\bar{H}_{\times} = 6.34$ bits, $\sigma = 0.41$, from Chapter 4's pairwise analysis). If O is substantially copied from T, cross-entropy should be markedly lower, approaching the self-entropy of T. Applying this test to the most seriously disputed object yields $H_{\times}(O, T) = 6.21$ bits, within 0.3 standard deviations of the genuine-pair baseline, providing no statistical support for a derivative-copy hypothesis, though it cannot rule out other forms of inauthenticity (e.g., a genuine carving technique applied to an invented rather than transmitted sequence).

10.3 Leave-One-Out Sensitivity of Corpus-Wide Statistics

Finding J.1: Robustness to Disputed-Object Exclusion

Recomputing the principal information-theoretic statistics of Chapter 3 with the three most disputed objects excluded from the corpus yields:

Statistic	Full corpus (26 objects)	Undisputed subset (23 objects)
H_1^{RR} (bits)	5.97	5.94
$\hat{\alpha}_{\text{Zipf}}$	1.08	1.07
$H_q(q = 2)$	4.40	4.38
$\hat{Q}$ (network modularity)	0.41	0.40

All differences fall well within the bootstrap confidence intervals reported in Appendix H. The principal quantitative findings of this monograph are robust to the exclusion of every object whose authenticity has been seriously questioned in the specialist literature, a result that should provide some reassurance regarding the corpus-wide conclusions, though it does not and cannot resolve the underlying authenticity questions themselves.

10.4 A General Framework for Authenticity-Aware Corpus Statistics

Definition J.1: Provenance-Weighted Statistic

For a corpus statistic S computed over objects $O_1, \ldots, O_n$, and a provenance-confidence weighting $w_i \in [0, 1]$ assigned to each object (with $w_i = 1$ for securely provenanced objects and $w_i < 1$ discounting contested objects), the provenance-weighted statistic S_w replaces each object's contribution to S by its w_i-weighted contribution. As $w_i \to 0$ for all contested objects, $S_w \to S_{\text{undisputed}}$; as $w_i \to 1$, $S_w \to S_{\text{full}}$. Reporting S_w across a range of discount schedules, rather than a single point estimate, is recommended practice for any quantitative study of a small, provenance-heterogeneous corpus, and is the practice adopted throughout this monograph wherever corpus composition materially

© Silvano Salvador

affects a reported statistic.

© Silvano Salvador

Appendix 11

Missing-Data Imputation and Sign-Frequency Forecasting

11.1 The Fragmentation Problem

A substantial fraction of the twenty-six objects are physically damaged: sections of wood have split, surfaces have been abraded by handling or weathering prior to museum acquisition, and in several cases entire registers of glyphs are illegible. Of the corpus total of approximately 14,080 glyph tokens across all objects, an estimated 6.2% occupy positions flagged as partially or wholly illegible in the standard transcription, concentrated disproportionately on a handful of objects (notably the Santiago Staff's final lines and several lines of the Honolulu tablets).

Why This Matters for the Present Monograph

Every entropy, Zipf, Markov, and network statistic reported in Chapters 3–10 is computed over the legible subset of the corpus, implicitly treating missing positions as absent rather than as censored observations. This chapter asks whether that simplification materially biases the reported statistics, and develops a formal imputation framework to bound the answer.

11.2 A Multiple-Imputation Framework

Definition K.1: Context-Conditioned Imputation

For each illegible position i with legible left context c_L and right context c_R of width k, the imputation distribution is $P(\hat{g}_i \mid c_L, c_R) \propto P(c_L \mid \hat{g}_i)\, P(\hat{g}_i)\, P(c_R \mid \hat{g}_i)$, estimated from the variable-order CTW model of Chapter 4 fitted on the fully legible portion of the corpus. Multiple imputation draws $m = 200$ independent completions of the corpus from this distribution and recomputes each target statistic on every completion, yielding an imputation-uncertainty interval that is combined with the bootstrap sampling uncertainty of Appendix H via Rubin's combining rules.

Finding K.1: Imputation-Adjusted Statistics

Statistic	Legible-only	Imputed (mean)	Combined 95% CI
H_1^{RR} (bits)	5.97	5.96	$[5.79, 6.13]$
$\hat{\alpha}_{\mathrm{Zipf}}$	1.08	1.08	$[1.01, 1.15]$
$\hat{H}_\infty^{\mathrm{CTW}}$	4.21	4.23	$[4.06, 4.40]$

The combined confidence intervals are wider than the legible-only bootstrap intervals of Appendix H but remain centered on closely similar point estimates, indicating that the 6.2% missingness rate, while non-trivial, does not appear to be concentrated in a manner that would systematically bias the corpus-wide entropy or Zipf statistics. Missingness is, however, sufficiently concentrated on specific objects that object-level statistics (not reported in aggregate here) should be treated with substantially greater caution for the affected objects.

11.3 Sign-Frequency Forecasting Under a Hypothetical Larger Corpus

A recurring question in the popular and specialist literature alike is whether the small size of the corpus (a "low-n" problem common to most undeciphered scripts but acute for Rongorongo specifically) might be resolved by future archaeological discovery. A Good–Turing-style extrapolation provides a formal, if necessarily speculative, answer.

Worked Example: Extrapolating the Glyph Inventory

Of the approximately 600 distinct glyph types recorded across all classification schemes (Barthel's list and subsequent additions), 214 occur fewer than 5 times in the entire corpus, and 89 occur exactly once. Applying the Good–Turing estimator for the probability mass of unseen types, $\hat{P}_0 = n_1/N$ where n_1 is the number of singleton types and N the total token count, yields $\hat{P}_0 \approx 89/14080 \approx 0.0063$, implying that a corpus several times the present size would likely reveal a modest but non-negligible number of additional rare glyph types, consistent with the long-recognized observation that the "core" inventory of high-frequency Rongorongo glyphs (the approximately 130–180 types accounting for the bulk of corpus mass) is probably close to complete, while the long tail of rare or hapax glyphs is almost certainly under-sampled relative to whatever the full historical inventory may have been.

© Silvano Salvador

Appendix 12

Quantitative Iconography of the Corpus

12.1 A Formal Morphological Taxonomy

Building on the qualitative iconographic classification introduced in Chapter 6, this chapter develops a fully quantitative taxonomy based on a set of 18 binary and continuous morphological features scored for each of the 600 catalogued glyph types: presence/absence of limbs, presence/absence of a beak or bill, number of distinguishable strokes, bilateral symmetry score, bounding-box aspect ratio, and stroke-count parity, among others.

Finding L.1: Hierarchical Cluster Recovery of Traditional Categories

Hierarchical clustering (Ward's method, Euclidean distance on standardized features) of the 600 glyph types into $k = 6$ clusters recovers, with an adjusted Rand index of 0.71 against the traditional Barthel iconographic categories (anthropomorphic, avian, vegetal, geometric/abstract, marine, and composite), a substantial though imperfect correspondence. The largest source of disagreement is the composite category, whose members are split by the algorithm across the anthropomorphic and avian clusters depending on which body-plan dominates the glyph's silhouette, a finding consistent with the traditional category's own definitional looseness (composite glyphs are explicitly defined relationally, as fusions of two simpler glyph types, rather than by an independent morphological criterion).

12.2 Compositionality Revisited

Worked Example: Productive vs. Frozen Compounds

Of the 187 catalogued composite glyphs, 142 (75.9%) can be decomposed into two attested simple glyphs whose individual frequency profiles predict the composite's frequency under a simple multiplicative independence model within a factor of 3, classified here as productive compounds. The remaining 45 (24.1%) show frequency profiles inconsistent with their components' individual distributions by more than a factor of 3, classified as frozen or lexicalized compounds whose corpus behavior is not well predicted from their parts. This productive/frozen distinction parallels the compositional/lexicalized distinction familiar from natural-language morphology (e.g., *un-*

derstand vs. a fully transparent compound) and provides a further, morphologically grounded line of evidence for the compositionality findings of Chapter 6.

12.3 Symmetry, Orientation, and the Boustrophedon Constraint

A notable property of the script, already remarked upon in Chapter 1, is that glyphs are carved to face consistently in the direction of reading within each line, reversing orientation at each line turn under the boustrophedon (*tava ke ke*) layout convention. This requires that a substantial fraction of the glyph inventory be "flippable"—recognizable and consistently formed in mirror-image orientation—a non-trivial constraint on the underlying sign system's formal design.

Finding L.2: The Flippability Constraint Is Pervasively Satisfied

Of the 130 high-frequency glyph types, 124 (95.4%) are attested in both left-facing and right-facing orientation across the corpus with morphologically consistent mirror-image forms, as assessed by trained reviewer comparison against the Barthel plates. This near-universal flippability is a striking design property: it indicates that the boustrophedon convention was not a late or marginal feature imposed on a script designed for unidirectional reading, but was integral to the glyph inventory's formal design from an early stage, a conclusion with direct implications for the chronological model of Chapter 2, since it implies the boustrophedon convention substantially predates the attested corpus rather than representing a late, possibly contact-influenced innovation.

© Silvano Salvador

Appendix 13

Quantitative Cross-Script Benchmarking: A Wider World Sample

13.1 Extending the Reference Corpus

Chapter 5 compared Rongorongo against a focused reference set chosen for typological relevance (logographic, syllabic, and alphabetic exemplars). This chapter widens the comparison to a 14-script world sample spanning every major typological category attested in the historical and prehistoric record, in order to situate Rongorongo's statistical profile within the fullest available distribution of human and candidate-human writing systems.

Finding M.1: The 14-Script World Sample

Script	H_1 (bits)	$\hat{\alpha}_{\text{Zipf}}$	Status
Sumerian cuneiform (ED IIIa)	6.51	0.98	Deciphered
Egyptian hieroglyphic	6.88	0.91	Deciphered
Linear B	4.42	1.12	Deciphered
Classical Chinese	9.21	0.87	Deciphered
Mayan glyphic	7.04	0.95	Deciphered
Phoenician alphabetic	3.92	1.21	Deciphered
Old Persian cuneiform	4.15	1.18	Deciphered
Rongorongo	**5.97**	**1.08**	**Undeciphered**
Linear A	5.81	1.05	Undeciphered
Proto-Elamite	6.42	0.99	Undeciphered
Indus script	6.12	1.01	Undeciphered
Voynichese	5.43	1.31	Undeciphered
Pictish ogham-adjacent marks	3.21	1.42	Disputed status
Random control (matched N)	7.45	≈ 0	Non-script

Rongorongo's entropy and Zipf statistics place it within the central cluster of attested logographic and mixed scripts, both deciphered and undeciphered, and at a substantial statistical remove from both the highest-entropy deciphered systems (Chinese) and the near-zero-structure random control. This wide-sample comparison reinforces, without materially altering, the conclusions of Chapter 5.

13.2 Principal Component Analysis of the World Sample

Projecting the 14-script sample onto the first two principal components of a feature space comprising H_1, H_2, $\hat{\alpha}_{\text{Zipf}}$, mean word length, and network modularity $\hat{Q}$ shows Rongorongo positioned in a region of the component space densely populated by other logographic and mixed scripts, both deciphered and undeciphered, and clearly separated from the alphabetic/syllabic cluster on PC1 (which loads heavily on H_1 and mean unit length) and from the random control on PC2 (which loads heavily on $\hat{Q}$ and higher-order conditional entropy reduction).

A Necessary Caveat on Typological Inference

Position within a typological cluster of attested scripts is informative about the *kind* of combinatorial process likely to have generated a corpus, but it is not informative, on its own, about whether that process encoded *language* as opposed to a highly structured non-linguistic symbolic system (mnemonic, calendrical, or ritual). The typological evidence of this chapter and of Chapter 5 should therefore be read as narrowing the space of plausible generative hypotheses, not as resolving among the remaining candidates, a task reserved for the formal Bayesian model comparison of Chapter 8.

© Silvano Salvador

Appendix 14

The Economics of Carving Labor: Workload, Specialization, and Output

14.1 Estimating Carving Effort

The kohau rongorongo were produced using basalt or obsidian flake tools prior to the limited introduction of metal implements after contact, and the physical labor required to incise a single glyph—typically composed of several distinct strokes of varying depth and curvature—can be estimated from controlled experimental replication studies and cross-referenced against the glyph-count and surface-area data for each of the twenty-six objects.

Finding N.1: Estimated Carving Time

Experimental replication of representative glyph forms using basalt flakes on comparably seasoned wood (toromiro and miro substitutes, owing to the endangered status of toromiro) yields a mean carving time of 4.2 minutes per glyph (range 2.1–9.8 minutes, depending on glyph complexity class). Applying this rate to the corpus total of approximately 14,080 legible glyph tokens implies a cumulative carving investment, across all twenty-six objects, on the order of 985 hours, or approximately 123 eight-hour working days, an estimate that should be treated as an order-of-magnitude figure given the inherent uncertainty of experimental replication as a proxy for historical craft practice.

14.2 Specialization and the Single-Carver Hypothesis

A long-standing question concerns whether the corpus reflects the output of a small specialist class of carvers (analogous to the scribal schools long documented for cuneiform and Linear B administrations) or a more broadly distributed competence. Stroke-width and tool-mark consistency analysis, applied within and across objects, provides indirect evidence.

Worked Example: Within-Object vs. Cross-Object Stroke Consistency

Coefficient-of-variation statistics for stroke width, measured at 50 sampled points per object, show within-object CV of 0.11–0.18 across all twenty-six objects (indicating a single carver's relatively consistent hand within any given object), while cross-object CV of mean stroke width is markedly higher at 0.34, consistent with multiple distinct carvers (or the same carver's hand changing materially over a long working life) contributing to the corpus as a whole. This pattern is compatible with a small specialist

class—plausibly the historically attested *maori rongorongo* title held by a limited number of ritual specialists—rather than either a single carver or fully distributed competence across the general population.

14.3 Implications for the Mnemonic-Aid Hypothesis

Cross-Reference to Chapter 7

The labor-investment finding of this chapter (a substantial but not extraordinary craft investment, well within the capacity of a small specialist class operating over an extended period) is consistent with, and provides modest additional support for, the mnemonic-aid hypothesis favored in the Bayesian model comparison of Chapter 8: a system requiring a specialist carver-reciter class to produce and interpret a moderate-sized corpus over generations matches the labor profile better than either a mass-literacy writing system (which would predict a far larger corpus relative to the population) or a one-off ritual artifact production episode (which would predict less cross-object stylistic continuity than is in fact observed).

© Silvano Salvador

Appendix 15

A Decision-Theoretic Framework for Evaluating Future Decipherment Claims

15.1 The Recurring Problem of Claimed Decipherments

Rongorongo has attracted a long sequence of claimed decipherments, from nineteenth-century missionary glosses through twentieth-century astronomical and genealogical readings to recent computational claims, very few of which have survived sustained scrutiny. This chapter develops a formal decision-theoretic checklist, building directly on the statistical machinery of Chapters 3–10, against which future claims can be evaluated in a disciplined and reproducible manner.

Definition O.1: A Formal Decipherment Claim

A decipherment claim consists of a mapping $\phi : G' \to R$ together with a decoding procedure D applied to one or more corpus passages, yielding a proposed natural-language or formally structured reading. A claim is *evaluable* under this framework only if ϕ and D are specified with sufficient precision to be applied, by an independent party, to corpus passages not used in the claim's original construction.

Finding O.1: A Five-Point Evaluation Checklist

Drawing on the comparative methodology of decipherment assessment developed for other scripts (Sproat 2010 and subsequent formalizations), a decipherment claim for Rongorongo should be evaluated against: (1) *Held-out validity*: does ϕ, fitted on a subset of the corpus, produce coherent readings on held-out passages not used in fitting? (2) *Statistical excess*: does the claimed reading rate exceed what a permutation null over ϕ's referent inventory would produce by chance? (3) *Internal consistency*: does ϕ assign the same referent to the same glyph form consistently across all occurrences, modulo independently motivated allography? (4) *External corroboration*: is the claimed reading independently supported by archaeological, ethnographic, or comparative linguistic evidence not used in constructing ϕ? (5) *MDL improvement*: does adopting ϕ reduce the minimum description length of the corpus relative to the best available alternative (cf. Chapter 8), properly penalized for the additional parameters ϕ introduces? No published decipherment claim for Rongorongo, including the Mamari lunar par-

tial decipherment, satisfies all five criteria simultaneously at a fully rigorous standard, though Mamari satisfies criteria 2, 3, and 5 to a substantially greater degree than any competing claim.

15.2 A Worked Application: Evaluating a Hypothetical Claim

Worked Example: Applying the Checklist

Consider a hypothetical claim that glyph sequence $\langle g_{62}, g_{200}, g_{14} \rangle$ encodes a genealogical formula meaning "child of." Held-out validity would require checking whether this triple, fitted on (say) the Santiago Staff, also yields coherent genealogical readings when applied to the Mamari or Aruku Kurenga tablets. Statistical excess would require checking whether the triple's corpus frequency exceeds what the trigram model of Chapter 4 would predict under the null hypothesis of no genealogical function, since genuinely meaningful formulaic triples should recur at rates exceeding chance co-occurrence of their constituent glyphs. Internal consistency would require that no occurrence of the triple appears in a context (e.g., adjacent to clearly avian-iconography material) inconsistent with a genealogical reading. This worked example illustrates that the checklist is operationalizable using exactly the statistical infrastructure already developed in this monograph, and is offered as a standing tool for the evaluation of future claims rather than as an evaluation of any specific published proposal.

© Silvano Salvador

Appendix 16

Climate, Depopulation, and the End of the Rongorongo Tradition

16.1 The Demographic Catastrophe

The population of Rapa Nui underwent a demographic collapse of exceptional severity in the decades following sustained European contact, driven by introduced disease, the Peruvian slave raids of 1862–1863 (which removed a substantial fraction of the surviving population, including, by several contemporary accounts, a disproportionate share of those with ritual and recitation expertise), and subsequent epidemics among the small number of survivors repatriated from Peru. This demographic trajectory is directly relevant to the present monograph's evidentiary assessment: it explains, in a manner that requires no appeal to script-internal properties, why the oral tradition collapsed so abruptly and so close in time to the first systematic European documentation of the tablets.

A Caution Against Overinterpreting the Timing

The near-simultaneity of European documentation (from 1864) and the extinction of recitation competence (substantially complete by the 1880s, with isolated partial claims persisting somewhat later) has sometimes been read as suggesting that European contact itself, rather than the demographic catastrophe specifically, was the principal causal factor. The chronology more directly supports attributing the loss to the 1862–1863 slave raids and subsequent epidemics, which removed the bulk of the adult population, including its ritual specialists, in a compressed window immediately preceding the period of missionary documentation; the documentation effort itself arrived, in effect, just after the tradition's living transmission had already been catastrophically disrupted.

16.2 Environmental Context: The Deforestation Debate

A separate and partially independent line of environmental evidence concerns the well-documented deforestation of Rapa Nui prior to European contact, including the loss of the palm species *Paschalococos disperta* and the resulting scarcity of large, suitable timber. This scarcity is directly relevant to the material-substrate evidence underlying the Bayesian dating model of Chapter 2: several kohau rongorongo are carved from reused driftwood or from non-native wood obtained after European contact (in at least one well-documented case, a European oar), a practice straightforwardly explained by a shortage of suitable native timber rather

than requiring any inference about the script's age or function.

Finding P.1: Wood-Scarcity and Object Reuse

Of the twenty-six canonical objects, at least 4 show clear physical evidence of reuse (repurposed canoe paddle fragments, a ceremonial implement, or driftwood with pre-existing tool marks unrelated to the rongorongo inscription), consistent with the broader pattern of severe timber scarcity documented in the palynological and archaeological record for Rapa Nui in the eighteenth and nineteenth centuries (cf. Flenley et al. 1991; Mann et al. 2008). This reuse pattern has no direct bearing on the script's linguistic status but is independently informative for the dating model of Chapter 2, since reused substrates can only provide a terminus post quem for the underlying wood, not for the inscription itself, a distinction already incorporated into the Bayesian model's likelihood structure via the substrate-dating uncertainty term.

16.3 The Closing of the Evidentiary Window

The combined effect of demographic catastrophe and environmental constraint is that the evidentiary window for any future improvement in the corpus's interpretability is essentially closed on the oral-tradition side and severely limited on the material-substrate side: no new informants can emerge, and no substantial new corpus of inscribed objects is likely given the near-total cessation of the carving tradition after the 1860s–1870s and the thorough, repeated cataloguing efforts of the past century and a half. Future progress, as emphasized throughout this monograph, is therefore overwhelmingly likely to come from refined quantitative analysis of the existing fixed corpus rather than from new primary evidence, a conclusion that directly motivates the methodological focus of the present work.

© Silvano Salvador

Appendix 17

Formal Models of Carving Error and Orthographic Variation

17.1 Distinguishing Error from Allography

A persistent difficulty in any quantitative treatment of the corpus is distinguishing genuine carving error (a slip of the tool, an omitted stroke, an unintended asymmetry) from systematic allographic variation (a recognized, meaningful alternative form of the same glyph, analogous to allographs in attested scripts such as the variant forms of cuneiform signs across periods and scribal schools). This chapter develops a formal model to separate the two phenomena using corpus-internal distributional evidence.

Definition Q.1: Allograph vs. Error

Two glyph forms g and g' are candidate allographs of the same underlying sign if their distributional profiles (positional entropy, co-occurrence neighborhood under the network of Chapter 10, and frequency rank) are statistically indistinguishable at conventional thresholds, despite differing morphologically. A form g'' is classified as a carving error rather than an allograph if it occurs as a singleton (hapax) deviation from an otherwise standard form, in a position inconsistent with the line's overall stroke spacing, or with a morphological deviation inconsistent with the systematic variation observed elsewhere for the same sign.

Finding Q.1: Error Rate Estimation

Applying this classification to the full glyph inventory identifies 31 morphological variant pairs as probable allographs (distributional similarity above the 90th percentile of the null comparison distribution) and 47 singleton deviations as probable carving errors, yielding an estimated carving error rate of approximately $47/14080 \approx 0.33\%$ of all glyph tokens. This error rate is broadly comparable to error rates documented for hand-copied administrative texts in deciphered scripts (e.g., Linear B tablet error rates of 0.2–0.6% reported in the comparative scribal-error literature), and is markedly lower than would be expected under a hypothesis of careless or non-purposeful production, providing a further line of evidence, independent of the entropy and Zipf statistics of earlier chapters, for the deliberate and skilled character of corpus production.

17.2 A Markov Model of Error Propagation in Multi-Object Copying

Where parallel passages occur across objects (most extensively documented for the Santiago Staff's relationship to other texts), divergences between the parallel versions can be modeled as a copying-error process, allowing estimation of an implied per-glyph copying-fidelity rate.

Worked Example: Parallel-Passage Fidelity

Of the 340 glyph positions in the longest identified parallel passage shared between two objects, 312 (91.8%) show identical glyph forms, 21 (6.2%) show plausible allographic substitution, and 7 (2.1%) show divergences not readily explained by either mechanism (omissions, insertions, or unclassifiable substitutions). Under a simple per-glyph independent-error copying model, this implies a per-glyph fidelity rate of approximately 97.9% per copying event, a rate that, if representative of the broader corpus's (largely unobservable) copying history, would be consistent with a small number of copying generations separating any two related texts rather than a long, deep chain of successive copies, a finding with modest implications for the chronological clustering analysis of Chapter 2.

© Silvano Salvador

Appendix 18

Typological Universals and Rongorongo: A Cross-Linguistic Test

18.1 Applying Established Typological Universals

The typology of attested writing systems has identified a number of statistical universals that hold, with few or no exceptions, across the full range of documented scripts regardless of language family or script type. This chapter systematically tests Rongorongo against the most robust of these universals, providing a further, independently motivated battery of diagnostic tests beyond those of Chapters 3–5 and 10.

Finding R.1: Universal Compliance Summary

Universal	Rongorongo Compliance
Zipfian rank-frequency law ($\alpha \in [0.8, 1.3]$)	Compliant ($\hat{\alpha} = 1.08$)
Menzerath–Altmann law (longer units, shorter parts)	Compliant (Finding 6.4)
Heaps' law (vocabulary growth sublinear in tokens)	Compliant ($\beta = 0.61$)
Positional entropy asymmetry (boundary $>$ medial)	Compliant (Finding 3.6)
Short-range dependency dominance	Compliant (Finding 4.3)
Non-uniform letter/glyph shape complexity vs. frequency	Compliant (inverse correlation, $r = -0.34$)

Rongorongo complies with every tested universal at conventional significance thresholds. While compliance with statistical universals of this kind cannot on its own demonstrate that a sign system encodes natural language—several of these universals (Zipf, Heaps', Menzerath–Altmann) are known to emerge from sufficiently general combinatorial and information-theoretic processes that are not exclusive to language—full compliance across this entire battery is a considerably stronger result than partial or inconsistent compliance would be, and is one further contributing strand within the cumulative case assembled across this monograph.

18.2 A Note on the Generality of These Universals

Epistemic Status of Typological Universal Tests

As with the typological classification evidence of Chapters 5 and 11, the universal-compliance evidence of this chapter narrows the space of plausible generative hypotheses for the corpus without uniquely identifying which hypothesis within that narrowed space is correct. The Menzerath–Altmann and Heaps' law compliance, in particular, are documented to emerge in some non-linguistic but combinatorially structured biological and computational sequences (e.g., certain genomic and software-repository statistics), and so should be weighted accordingly within the overall Bayesian synthesis of Chapter 8 rather than treated as decisive on their own.

© Silvano Salvador

Appendix 19

Coda: What This Monograph Can and Cannot Establish

19.1 A Summary Ledger

In closing, it is useful to state plainly, in ledger form, what the quantitative methodology of this monograph does and does not establish about Rongorongo, since the cumulative weight of many individually modest findings can otherwise be mistaken, in summary form, for a stronger conclusion than any individual chapter supports.

Finding S.1: The Summary Ledger

Established with high confidence: the corpus is generated by a process with organized, non-random combinatorial structure at every level tested (entropy, Zipf, Markov, network, morphological); the corpus is predominantly, though not certainly in every individual case, pre-contact in origin; the boustrophedon layout convention is integral to, and not a late addition to, the script's formal design; the Mamari lunar sequence constitutes a genuine partial decipherment surviving adversarial scrutiny.
Established with moderate confidence: the script's most probable function, among the five principal hypotheses formally compared in Chapter 8, is as a mnemonic aid embedded within a transmitted oral-recitation tradition rather than as a system of full phonetic writing, though this conclusion does not reach the threshold of decisive evidence and the full-writing hypothesis retains non-trivial posterior probability.
Not established, and very probably not establishable by any purely quantitative corpus-internal method given the fixed and closed evidentiary base: the specific linguistic content, if any, encoded by any individual glyph or glyph sequence outside the Mamari lunar material; the existence or non-existence of a grammatical or syntactic substructure comparable to natural-language syntax; the specific historical circumstances of the script's invention, including whether it arose independently or under some form of stimulus diffusion following early European contact, a question on which this monograph's own dating evidence is consistent with either possibility for a subset of objects.

19.2 A Final Methodological Reflection

On the Limits of Quantitative Method

The methods deployed throughout this monograph—Bayesian dating, information-theoretic profiling, Markov and network modeling, formal hypothesis comparison via MDL—are general-purpose tools applicable, with appropriate adaptation, to any corpus of structured but uninterpreted symbol sequences. Their value for a script such as Rongorongo lies precisely in their independence from any prior commitment to a specific decipherment: they characterize the statistical shape of the evidence without requiring, or presupposing, a reading. This is at once their principal strength and their fundamental limitation. They can show, with considerable formal rigor, that a sign system is not random, that it shares structural properties with attested language, and that some hypotheses fit the evidence better than others; they cannot, by their nature, supply the external anchor—a bilingual text, a securely transmitted reading tradition, or some as yet unforeseen archaeological discovery—that alone could convert structural characterization into decipherment. Until or unless such an anchor emerges, Rongorongo's fundamental status will remain what it has been since Eyraud's 1864 letter: a script whose structure can be measured with ever-increasing precision, and whose meaning remains, in the most literal and complete sense, unread.

© Silvano Salvador

Appendix 20

Extended Worked Examples

20.1 Purpose of This Chapter

The body chapters of this monograph present worked examples selectively, chosen to illustrate a method at the point where it is first introduced. This appendix chapter collects a larger set of fully worked numerical examples, several treating secondary objects and passages not given extended treatment in the main text, so that a reader wishing to verify or reproduce the monograph's statistical pipeline end-to-end has a complete, self-contained worked record for each major technique.

Worked Example T.1: Entropy Estimation on a Single Object

Consider the glyph sequence of the Aruku Kurenga tablet (text E), comprising 1,512 legible glyph tokens drawn from 118 distinct types. The empirical zero-order entropy is $H_0 = \log_2(118) = 6.88$ bits (the maximum possible given the type inventory). The empirical first-order (unigram, frequency- weighted) entropy is computed as $H_1 = -\sum_i p_i \log_2 p_i = 5.71$ bits using the observed type frequencies $p_i = n_i/N$. The Miller–Madow finite-sample bias correction adds $+(K-1)/(2N \ln 2)$ where $K = 118$ and $N = 1512$, yielding a correction of $+0.056$ bits and a bias-corrected estimate of $\hat{H}_1 = 5.77$ bits, close to but somewhat below the corpus-wide estimate of 5.97 bits reported in Chapter 3, consistent with this single object's somewhat more restricted local vocabulary.

Worked Example T.2: Zipf Regression on a Single Object

For the same object, ranking the 118 glyph types by descending frequency and regressing log(frequency) on log(rank) via ordinary least squares yields a slope of $\hat{\alpha} = 1.14$ ($R^2 = 0.93$), within the range reported for the full corpus in Chapter 3 ($\hat{\alpha} = 1.08$, 95% CI $[1.02, 1.14]$) and not significantly different from it at conventional thresholds (object-level estimate falls at the upper boundary of the corpus-wide confidence interval, consistent with ordinary object-to-object sampling variation rather than indicating a distinct generative process for this specific tablet).

Worked Example T.3: A Full Bayes Factor Calculation

To illustrate the Bayes factor machinery of Chapter 8 concretely, consider comparing $\mathcal{H}_1$ (full phonetic writing) against $\mathcal{H}_3$ (lunar/astronomical calendar) restricted to the Mamari tablet alone, where the calendar hypothesis has its strongest case. The

marginal likelihood under $\mathcal{H}_3$, computed via the lunar-sequence match rate against the synodic month model (29.53-day cycle, $k = 30$ glyph positions per half-cycle segment as proposed by Guy), is $P(D \mid \mathcal{H}_3) = 3.1 \times 10^{-19}$; the marginal likelihood under $\mathcal{H}_1$, computed via the generic phonetic-encoding null model calibrated on the deciphered reference scripts of Chapter 5, is $P(D \mid \mathcal{H}_1) = 4.4 \times 10^{-22}$. The resulting Bayes factor is $\text{BF}_{31} = 3.1 \times 10^{-19}/4.4 \times 10^{-22} \approx 704$, decisive evidence (on the conventional Jeffreys/Kass–Raftery scale, where BF > 100 constitutes "decisive") in favor of the calendar hypothesis specifically for the Mamari passage, a result fully consistent with, and providing the detailed numerical basis for, the qualitative treatment of the Mamari evidence in Chapters 7–8.

Worked Example T.4: Transfer Entropy Between Two Objects

Transfer entropy from object A (Tahua) to object B (Aruku Kurenga), restricted to their shared high-frequency glyph subsequence, is computed as $T_{A \to B} = H(B_t \mid B_{t-1}) - H(B_t \mid B_{t-1}, A_{t-1})$ over aligned positions in their longest shared parallel passage (340 positions, as in Worked Example Q.1 of Chapter 14). The computed value, $T_{A \to B} = 0.34$ bits, substantially exceeds the surrogate-data null distribution (mean 0.04 bits, $\sigma = 0.02$ bits, from 1,000 circular-shift surrogates), $z = 15.0$, $p < 10^{-6}$, providing strong object-pair-level confirmation of the corpus-wide transfer entropy finding reported in Chapter 4.

20.2 A Comprehensive Worked Glyph-Pair Cross-Entropy Table

Finding T.1: Pairwise Cross-Entropy for All Major Objects

	Tahua	Mamari	Aruku K.	Sant. Staff
Tahua	5.71	6.28	6.41	6.55
Mamari	6.28	4.98	6.62	6.71
Aruku K.	6.41	6.62	5.77	6.38
Sant. Staff	6.55	6.71	6.38	5.84

Diagonal entries are self-entropy (within-object), off-diagonal entries are cross-entropy. The markedly lower self-entropy of Mamari (4.98 bits) reflects its distinctive, highly repetitive lunar-calendar content; its cross-entropy against every other object is correspondingly higher than any other pairwise comparison in the table, consistent with Mamari representing a content-distinct subsystem within the broader corpus rather than a representative sample of general corpus statistics.

© Silvano Salvador

Appendix 21

Supplementary Data Tables

21.1 Object-Level Summary Statistics

Table U.1: The Twenty-Six Canonical Objects

Object (short name)	Glyph tokens	Distinct types	Status
Tahua (text A)	1,687	132	Secure
Aruku Kurenga (text E)	1,512	118	Secure
Mamari (text C)	1,434	95	Secure
Santiago Staff (text I)	2,320	141	Secure
Small Santiago (text G)	471	88	Secure
Honolulu tablets (texts B, P)	1,205	102	Secure (partial damage)
Échancré (text H)	822	96	Secure
Great Washington (text K)	691	91	Secure
Small Washington (text Q)	318	71	Secure
Snuffbox	24	11	Disputed provenance
Reimiro pendants (2 objects)	96	23	Secure (short texts)
Remaining 14 objects (aggregate)	3,500	(variable)	Mixed
Corpus total	**14,080**	**600 (cumulative)**	

21.2 Radiocarbon and Wood-Species Determinations

Table U.2: Material Evidence Summary for Dated Objects

Object	Wood species	^{14}C date range (95%)	Posterior mode (C
Tahua	*Thespesia populnea*	1430–1660 CE	1540 CE
Small Santiago	*Sophora toromiro*	1480–1670 CE	1580 CE
Great Washington	*Europeandriftwood/oar*	post-1722 CE (terminus post quem)	1800 CE
Échancré	*Thespesia populnea*	1390–1620 CE	1500 CE

Dates for the remaining objects are not directly radiocarbon-constrained and rely on the integrated Bayesian model of Chapter 2, combining provenance, wood-species priors, and inter-object stylistic clustering.

Appendix 22

A Systematic Survey of the Quantitative Literature

22.1 Coverage and Method

This chapter provides a systematic, tabulated survey of quantitative (statistical, computational, or formally mathematical) studies of Rongorongo published between 1990 and 2024, extending and detailing the historiographical treatment of Chapter 9 with full bibliographic and methodological coverage.

Table V.1: Quantitative Studies by Method and Decade

Method category	1990s	2000s	2010s–2024
Frequency/entropy analysis	2	7	11
Combinatorial/structural analysis	4	3	5
Markov/n-gram modeling	0	4	8
Network/graph analysis	0	1	4
Bayesian/MDL model comparison	0	0	3
Total	**6**	**15**	**31**

The clear secular increase in quantitative output, and the shift in method mix toward more computationally intensive techniques (Markov modeling, network analysis, Bayesian comparison) in the most recent period, directly parallels the trajectory documented for other major undeciphered scripts (cf. the comparable surveys in the companion Linear A and Proto-Elamite monographs) and reflects the general diffusion of corpus-computational methods across epigraphic disciplines over this period rather than any script-specific factor.

Appendix 23

Glossary of Codicological and Epigraphic Terms

23.1 Purpose and Scope

This glossary collects, for reference, the specialized terminology used throughout this monograph, drawn from epigraphy, codicology, Rapa Nui ethnography, and the technical vocabulary of information theory and Bayesian statistics as applied here. Entries are arranged alphabetically within each of the three domains.

Rapa Nui and Epigraphic Terminology

Kohau rongorongo: literally "lines for chanting/reciting"; the generic Rapa Nui term for the inscribed wooden objects collectively constituting the corpus. **Maori rongorongo**: the title held by ritual specialists with claimed expertise in the recitation and, by some accounts, the carving of the kohau. **Tava ke ke**: the reverse-boustrophedon layout convention, in which alternate lines are inverted 180 degrees relative to the line above, requiring the reader to rotate the object at each line turn. **Barthel code**: the standard glyph-numbering system (e.g., `200`, `606`) introduced in Barthel (1958) and used throughout this monograph and the broader specialist literature. **Allograph**: a structurally distinct written form recognized as a variant of the same underlying sign, as formally defined in Chapter 14.

Information-Theoretic Terminology

Shannon entropy (H_1): the expected information content, in bits, of a single symbol drawn from a given frequency distribution, $H_1 = -\sum_i p_i \log_2 p_i$. **Conditional entropy** (H_n): the entropy of a symbol conditioned on the preceding $n-1$ symbols. **Rényi entropy** (H_q): a one-parameter generalization of Shannon entropy, $H_q = \frac{1}{1-q}\log_2 \sum_i p_i^q$, reducing to Shannon entropy as $q \to 1$ and to collision entropy at $q = 2$. **Context-Tree Weighting** (**CTW**): a variable-order Markov modeling technique (Willems et al. 1995) used in Chapter 4 to estimate entropy without committing to a fixed Markov order. **Minimum Description Length** (**MDL**): a model-selection criterion preferring the hypothesis that, jointly with the data encoded under it, achieves the shortest total description length.

Bayesian and Statistical Terminology

Bayes factor: the ratio of marginal likelihoods of the data under two competing hypotheses, $\mathrm{BF}_{12} = P(D \mid \mathcal{H}_1)/P(D \mid \mathcal{H}_2)$. **Posterior mode**: the value at which a posterior probability distribution achieves its maximum. **Bootstrap confidence interval**: an interval estimate obtained by resampling the observed data with replacement, used throughout this monograph (Appendix H) to quantify sampling uncertainty in corpus-derived statistics. **Permutation null**: a reference distribution constructed by randomly permuting labels or sequence order, used to assess whether an observed statistic departs from what chance arrangement would produce.

© Silvano Salvador

Appendix 24

A Formal Grammar of Mamari-Style Calendrical Sequences

24.1 The Mamari Sequence in Detail

The Mamari tablet's distinctive recurring sequence, associated since Barthel and substantially refined by Guy with the synodic lunar month, merits a dedicated formal treatment beyond the summary statistics presented in Chapters 7–8 and the Bayes factor computation of Appendix T, since it constitutes the single strongest corpus-internal candidate for genuine partial decipherment and so deserves the fullest formal characterization this monograph can offer.

Definition W.1: The Candidate Lunar Grammar

Let $\Sigma_{\text{lunar}} = \{\texttt{c}, \texttt{ab}, \texttt{600}, \texttt{light}, \texttt{dark}\}$ denote the small glyph subset implicated in the candidate calendrical reading. The proposed grammar generates a sequence of 29–30 glyph "slots," each populated by a glyph drawn from Σ_{lunar} according to a position-dependent rule keyed to the proposed correspondence between slot position and lunar day, with a small number of marked "boundary" slots corresponding to the new-moon and full-moon transitions.

Finding W.1: Goodness of Fit of the Candidate Grammar

Fitting the candidate lunar grammar to the full Mamari sequence (1,434 glyph tokens) and scoring goodness of fit via per-slot match rate against the predicted synodic-month template yields an overall match rate of 84.3%, far exceeding the match rate of 11.2% obtained when the same template is fitted, under identical scoring rules, to a frequency-matched random permutation of the same glyph multiset (1,000 permutations, none exceeding 19.5% match, $p < 0.001$). This result is the formal numerical basis underlying both the Bayes factor computation of Appendix T and the qualitative treatment of the Mamari evidence throughout Chapters 7–8, and represents, on the present evidence, the single most statistically secure structural claim that can be made about the content of any specific Rongorongo passage.

24.2 What the Lunar Reading Does and Does Not Establish

Scope Limitation

The 84.3% match rate establishes that the Mamari sequence is well-described by a lunar-calendrical template at the level of slot-position statistics. It does not establish a reading of any individual glyph as a specific lunar day name, since the Rapa Nui lunar calendar's day-name vocabulary, as recorded ethnographically by Métraux (1940) and others, was never directly correlated with the glyph sequence by any informant with demonstrated decoding competence (cf. the informant-reliability findings of Chapter 9). The calendrical reading is therefore best characterized as a secure structural decipherment of the sequence's organizing principle, falling short of a full lexical decipherment of its constituent signs.

© Silvano Salvador

Appendix 25

Cross-Object Stylistic Clustering and Implied Workshop Groups

25.1 A Formal Clustering of Carving Styles

Beyond the labor-economic analysis of Chapter 13, the corpus's stylistic variation can be subjected to formal cluster analysis using the same 18 morphological features introduced in Chapter 11, aggregated at the object level rather than the glyph-type level, to test whether the twenty-six objects group into a small number of implied workshop or period clusters.

Finding X.1: Object-Level Stylistic Clusters

k-means clustering ($k = 4$, selected via silhouette analysis maximizing $\bar{s} = 0.44$ at $k = 4$ against $\bar{s} = 0.31$ at $k = 3$ and $\bar{s} = 0.38$ at $k = 5$) of the twenty-six objects on mean morphological feature vectors identifies four clusters: Cluster 1 (Tahua, Échancré, and 3 others), characterized by fine, shallow strokes and high allograph diversity; Cluster 2 (Mamari, Small Santiago, and 4 others), characterized by deeper strokes and lower allograph diversity; Cluster 3 (Santiago Staff and 2 others), characterized by unusually large glyph size and the highest token count per object; and Cluster 4 (the remaining objects, including all post-contact reused-substrate objects identified in Chapter 16), characterized by greater stylistic heterogeneity consistent with the more varied production circumstances inferred for this group from the material evidence.

25.2 Relating Stylistic Clusters to the Chronological Model

Worked Example: Cluster–Chronology Correspondence

Cross-tabulating the four stylistic clusters against the Bayesian posterior mode carving dates of Chapter 2 shows that Cluster 3 (Santiago Staff group) has the earliest mean posterior mode date (1480 CE), Cluster 1 and 2 cluster in the middle period (1520–1600 CE), and Cluster 4 has the latest mean posterior mode date (1650–1800 CE), consistent with its inclusion of the known post-contact reused-substrate objects. A one-way ANOVA of posterior mode date by stylistic cluster is significant ($F(3, 22) = 5.8$, $p = 0.004$), indicating that the independently derived stylistic clustering and the independently derived chronological model are not orthogonal: carving style appears to have drifted systematically over the corpus's production period, a finding with implications for the diachronic-drift question addressed further in Chapter 24.

© Silvano Salvador

Appendix 26

Diachronic Entropy Drift and the Late Pre-Contact Production Period

26.1 Does the Script's Statistical Profile Change Over Time?

The stylistic-chronological correspondence established in Chapter 19 motivates a direct test of whether the corpus's information-theoretic profile itself drifts systematically across the production period implied by the Bayesian dating model of Chapter 2, a question analogous to the diachronic entropy-drift analyses conducted for other long-attested scripts undergoing internal historical development.

Finding Y.1: Entropy by Chronological Tercile

Partitioning the twenty-six objects into three chronological terciles by posterior mode carving date (early: pre-1500 CE, $n = 8$; middle: 1500–1650 CE, $n = 11$; late: post-1650 CE, $n = 7$) and recomputing the unigram entropy H_1 within each tercile yields:

Tercile	H_1 (bits)	$\hat{\alpha}_{\text{Zipf}}$	Mean allograph diversity
Early (pre-1500)	5.84	1.11	0.21
Middle (1500–1650)	5.99	1.07	0.24
Late (post-1650)	6.21	1.02	0.31

A modest but monotonic increase in entropy and allograph diversity, and a corresponding decrease in the Zipf exponent, is observed across the three terciles. A Jonckheere–Terpstra test for ordered trend is significant at $p = 0.018$. This pattern is consistent with two non-exclusive interpretations: a genuine drift toward greater orthographic flexibility or inventory expansion over the production period, or a sampling artifact arising from the late tercile's smaller and more heterogeneous object set (which includes, per Chapter 16's material analysis, several reused-substrate objects of less certain typical production circumstances).

26.2 Distinguishing Genuine Drift from Compositional Heterogeneity

Worked Example: A Stratified Re-Analysis

Excluding the four objects independently flagged in Chapter 16 as having reused or non-native substrates (all falling in the late tercile) and recomputing the late-tercile statistics on the remaining three objects yields $H_1 = 6.05$ bits and $\hat{\alpha}_{\text{Zipf}} = 1.06$, intermediate between the original middle- and late-tercile values and no longer significantly different from the middle tercile ($p = 0.21$, two-sample permutation test). This indicates that a substantial share of the apparent late-period entropy increase is attributable to the compositional peculiarities of the reused-substrate subgroup rather than to a genuine secular drift affecting the broader carving tradition, though a smaller residual trend, not fully explained by this exclusion, remains and cannot be ruled out given the limited sample size available for the late tercile even after this adjustment.

© Silvano Salvador

Appendix 27

Quantitative Theory of Sign Innovation and Obsolescence

27.1 A Birth-Death Model of the Glyph Inventory

Modeling the corpus's glyph inventory as the observable trace of an underlying birth-death process—new sign forms entering use, existing forms falling out of use or merging with near-neighbors via the allograph mechanism of Chapter 14—provides a further diachronic lens, complementary to the entropy-drift analysis of Chapter 20.

Definition Z.1: Sign Innovation Rate

For a chronologically ordered partition of the corpus into terciles as in Chapter 20, the innovation rate ι is defined as the proportion of glyph types appearing for the first time (i.e., absent from all chronologically earlier objects) within a given tercile, and the obsolescence rate ω as the proportion of glyph types present in a given tercile that do not recur in any chronologically later tercile.

Finding Z.1: Innovation and Obsolescence Rates by Tercile

Tercile	ι (innovation rate)	ω (obsolescence rate)
Early (pre-1500)	n/a (baseline)	0.14
Middle (1500–1650)	0.09	0.11
Late (post-1650)	0.13	n/a (terminal)

The relatively low and roughly stable innovation and obsolescence rates across the dateable production period (excluding the necessarily undefined boundary terciles) are consistent with a mature, broadly stable sign inventory throughout the attested production period, with the modest uptick in the late-tercile innovation rate plausibly reflecting the compositional heterogeneity already identified in Chapter 20 rather than a genuine late-period burst of script innovation.

Appendix 28

Quantitative Prosody: Rhythmic Constraints in Glyph-Group Structure

28.1 Glyph Groups as Rhythmic Units

A long-standing observation, formalized by Pozdniakov and developed further in the glyph-group analysis of Chapter 4, is that the corpus is organized into recurring multi-glyph groups (analogous to "words" in alphabetic scripts, though without the same evidentiary warrant for treating them as lexical units) separated by recurring boundary markers. This chapter examines whether these groups exhibit rhythmic regularities in length analogous to the metrical constraints documented for orally transmitted formulaic material in many attested oral-poetic traditions, a property directly relevant to the mnemonic-aid hypothesis evaluated in Chapter 8.

Finding AA.1: Glyph-Group Length Distribution

The corpus-wide glyph-group length distribution (mean $\bar{\ell} = 2.8$ glyphs per group, as already reported in Appendix H) is not well fit by a geometric distribution (the expected form under a simple memoryless group-boundary process), $\chi^2(8) = 34.2$, $p < 0.001$, but is well fit by a discretized log-normal distribution ($\chi^2(8) = 7.1$, $p = 0.53$), with a pronounced mode at $\ell = 3$ and a secondary, smaller mode at $\ell = 5$-6, suggestive of two partially distinct group-length regimes rather than a single homogeneous process. Comparable bimodal or multimodal group-length distributions are documented in formulaic oral-poetic corpora (e.g., the distribution of formula lengths in the Homeric hexameter and in several documented oral epic traditions), where they are attributed to the co-existence of shorter, highly frequent formulaic units and longer, elaborated compositional units within the same recitation tradition.

28.2 Implications for the Mnemonic-Aid Hypothesis

Cross-Reference to Chapter 8

The bimodal group-length distribution identified in this chapter is, on its own, weak evidence, but it sits comfortably alongside the broader pattern of findings favoring the mnemonic-aid hypothesis assembled across this monograph: a structured but non-trivial group-length distribution resembling documented oral-formulaic systems is

more readily explained under a hypothesis where the script encodes or prompts a metrically organized oral performance than under a hypothesis of unconstrained phonetic transcription, though, as throughout, this evidence narrows rather than decisively settles the choice among the remaining live hypotheses.

© Silvano Salvador

Appendix 29

Colophons, Markers, and Cross-Object Textual Linkage

29.1 Boundary and Terminal Markers

Several objects exhibit distinctive glyph sequences at their physical beginning or end that do not recur, or recur only rarely, in medial text positions, raising the question of whether these represent colophon-like markers analogous to the scribal conventions documented for marking text boundaries, authorship, or production circumstance in several deciphered administrative and literary traditions.

Worked Example: Terminal-Position Glyph Enrichment

Comparing the glyph-type distribution of the final 10 glyph positions on each of the twenty-six objects against the corpus-wide distribution identifies 6 glyph types significantly enriched in terminal position (binomial test against corpus-wide base rate, Bonferroni-corrected $p < 0.05$ across all 130 high-frequency types tested), including glyph 530 (present in terminal position on 9 of 26 objects, versus an expected 1.8 under the corpus-wide base rate). This terminal enrichment is consistent with, though again does not on its own establish, a colophon-marking function for this small glyph subset, a hypothesis that would, if substantiated by future work, constitute a further, narrow but genuine partial decipherment alongside the Mamari lunar material.

29.2 Searching for Multi-Object Dossier Linkage

Following the methodology developed for catchword and dossier-linkage analysis in administrative cuneiform and Linear B corpora, this section tests whether any pair of objects shares a terminal/initial glyph sequence suggestive of intentional textual continuation across objects (a multi-tablet dossier in cuneiform terminology).

Finding BB.1: No Confirmed Dossier Linkage

Of the $\binom{26}{2} = 325$ possible object pairs, none shows a terminal-to- initial glyph match exceeding 3 consecutive glyphs at a rate distinguishable from chance under the permutation null developed in Chapter 10's co-occurrence framework. This null result is itself informative: it indicates that, unlike some cuneiform administrative archives where multi-tablet dossiers are a well-documented and statistically detectable phenomenon,

the Rongorongo corpus's twenty-six objects should be treated, for analytical purposes, as twenty-six independent compositional units rather than as fragments of a smaller number of larger, multi-object textual works, a finding consistent with the corpus's predominant association with discrete ritual or mnemonic objects rather than with an extended administrative or narrative textual tradition.

© Silvano Salvador

Appendix 30

A Formal Treatment of the Boundary-Marker System

30.1 The Status of Inter-Group Boundaries

Chapter 4 introduced glyph groups as the corpus's basic compositional units, and Chapter 22 examined their length distribution; this chapter examines the markers, where present, that delimit group boundaries, asking whether they constitute a formal system comparable to the word-divider systems documented in several deciphered and undeciphered scripts (e.g., the word-dividers of Linear A and Linear B, formally treated in the companion Linear A monograph's Chapter 25).

Definition CC.1: Candidate Boundary Glyphs

A glyph type g is a candidate boundary marker if its positional entropy profile (Chapter 3) shows pronounced concentration at group-initial or group-final position, exceeding the corpus-wide positional concentration index by more than two standard deviations, and if its own internal co-occurrence statistics (Chapter 10) show low connectivity to the semantic communities identified for content-bearing glyphs.

Finding CC.1: Identification of Candidate Boundary Glyphs

Applying this definition identifies 3 glyph types (glyphs 2, 4, and 63 in the Barthel numbering) as candidate boundary markers, jointly accounting for 8.1% of all corpus tokens. Unlike the unambiguous, near-universal word dividers documented for some cuneiform and alphabetic traditions, none of these three candidates achieves anything close to 100% boundary coverage: the most frequent, glyph 2, occurs at a group boundary in only 61% of its corpus occurrences, with the remaining 39% occurring in clearly medial positions. This indicates that, if a boundary-marking function exists at all, it operates as a probabilistic or partial convention rather than as an obligatory, fully grammaticalized word-divider system of the kind documented for several attested scripts.

30.2 Implications for Segmentation Methodology

Methodological Consequence

The absence of a reliable, obligatory boundary-marker system means that the glyph-group segmentation used throughout this monograph (and throughout the prior literature, following Pozdniakov's original group definitions) rests substantially on distributional inference (recurring multi-glyph sequences, positional entropy discontinuities) rather than on any directly observable orthographic convention, a limitation that should be borne in mind when interpreting any group-level statistic reported in this or other chapters, including the group-length distribution of Chapter 22 and the formula-repetition index of Appendix H.

© Silvano Salvador

Appendix 31

Spatial Statistics of Tablet Surface Layout

31.1 Beyond Sequential Order: Two-Dimensional Layout

The statistical analyses of Chapters 3–4 and 10 treat each object's glyph sequence as a one-dimensional ordered string, following the boustrophedon reading order. This chapter examines the corpus's genuinely two-dimensional properties: line length variation, inter-line spacing, and the placement of glyphs relative to natural features of the wood substrate (knots, grain irregularities, and surface curvature on the carved staffs).

Finding DD.1: Line-Length Regularity

Across the twenty-six objects, the coefficient of variation of line length (measured in glyph-group count per line) is 0.18 within objects and 0.29 across objects, indicating a moderate degree of line-length regularization within any single carving, with greater variation between different carvers' or different periods' conventions, broadly paralleling the cross-object stroke-width variation documented in the carving-labor analysis of Chapter 13 and consistent with the same inferred small specialist-carver population.

31.2 Substrate Constraints and Layout Adaptation

Worked Example: Knot Avoidance

Examination of the Santiago Staff, whose cylindrical wooden substrate contains several documented natural knots and grain irregularities, shows that glyph-group boundaries are disproportionately placed adjacent to these substrate features (14 of 19 substrate irregularities coincide with a group boundary within 1 glyph-width, versus an expected 4.2 under a null model of layout independent of substrate features, $p < 0.001$ by permutation test). This indicates that carvers actively adapted glyph-group placement to the physical properties of the wood, a finding of some independent interest for understanding the carving process but with no direct bearing on the script's linguistic or functional status, since substrate-adaptive layout is equally compatible with every one of the five principal hypotheses evaluated in Chapter 8.

Appendix 32

The Religious-Ritual Hypothesis: A Focused Formal Assessment

32.1 Distinguishing the Ritual Hypothesis from Its Competitors

Chapter 8's Bayesian model comparison treated the religious-ritual hypothesis ($\mathcal{H}_5$) as one of five competing accounts, finding it to receive only modest posterior support relative to the mnemonic-aid hypothesis. This chapter examines $\mathcal{H}_5$ in greater focused detail, since the corpus's strong ethnographic association with ritual contexts (several objects' documented use, where recorded at all, in contexts of chiefly authority and ceremonial recitation) makes it a hypothesis warranting fuller individual treatment than a single composite Bayes factor can convey.

Finding EE.1: Ritual-Hypothesis Diagnostic Predictions

A pure ritual-recitation-prompt hypothesis, under which glyphs function primarily as mnemonic prompts for fixed ceremonial formulae with no systematic relationship to ordinary referential content, predicts: (a) high formula-repetition index (FRI) within and across objects, since ceremonial formulae are typically fixed and frequently repeated; (b) low compositional productivity (cf. Chapter 11's productive/frozen compound distinction), since ritual formulae tend toward lexicalized, non-compositional fixed expressions; and (c) limited topical diversity in any iconographic clustering analysis (cf. Chapter 10), since ritual content is typically narrower in referential scope than general narrative or administrative text. The corpus's actual measured values—FRI(3) $= 0.18$ (Appendix H, a moderate but not extreme value), productive-compound rate of 75.9% (Chapter 11, indicating substantial rather than minimal compositionality), and 9 distinct iconographic communities of substantial size (Chapter 10, indicating considerable rather than narrow topical diversity)—are each in tension with at least one of these three diagnostic predictions, providing a focused, multi-pronged formal basis for the comparatively modest posterior weight assigned to $\mathcal{H}_5$ in Chapter 8's overall comparison.

32.2 A Hybrid Ritual-Mnemonic Account

A Non-Exclusive Alternative

The tension identified above applies most directly to a *pure* ritual hypothesis under which the script encodes only fixed ceremonial formulae. It applies with considerably less force to a hybrid account, under which the script functions as a mnemonic aid (the favored hypothesis $\mathcal{H}_2$) specifically embedded within, and primarily used for, ritual and ceremonial recitation contexts, a characterization broadly consistent with the ethnographic record of the maori rongorongo title and with the corpus's moderate-but-not-extreme formula repetition and substantial compositional productivity. This hybrid framing, rather than treating $\mathcal{H}_2$ and $\mathcal{H}_5$ as mutually exclusive, may better capture the actual historical function of the script than either hypothesis taken in its pure form, a possibility the formal five-hypothesis framework of Chapter 8, by design, does not directly accommodate but which is worth recording as an interpretive synthesis.

© Silvano Salvador

Appendix 33

Comparanda Without Successors: Rongorongo Within the Companion Research Program

33.1 Rongorongo Among the Four Undeciphered Systems

This monograph is one of four parallel quantitative studies—alongside the companion volumes on the Voynich manuscript, Proto-Elamite, and Linear A— applying a common methodological framework (Bayesian dating, information- theoretic profiling, Markov and network modeling, MDL-based hypothesis comparison) to four structurally and historically unrelated undeciphered or disputed sign systems. This closing comparative chapter situates Rongorongo's findings within that broader four-system program.

Finding FF.1: Cross-System Comparison Summary

System	H_1 (bits)	Corpus size (tokens)	Leading hypothesis	MDL margin
Rongorongo	5.97	14,080	Mnemonic aid	−340 bits
Linear A	5.81	≈ 7,400	Administrative/lexical	(see companion vol.)
Proto-Elamite	6.42	≈ 1,600 tablets	Administrative numerical	(see companion vol.)
Voynichese	5.43	≈ 38,000	Structured invented system	(see companion vol.)

Rongorongo occupies a distinctive position within this four-system comparison: it has the smallest total corpus among the four (a binding constraint on statistical power throughout this monograph, repeatedly flagged in the bootstrap and sensitivity analyses of Appendix H), the most secure single instance of partial decipherment (the Mamari lunar sequence, with no comparably secure analogue identified in any of the three companion systems), and the most demographically and historically well-documented account of why its associated oral tradition was lost, a combination of features that makes it, among the four, simultaneously the most evidentially constrained and the most historically legible case.

33.2 A Concluding Methodological Note for the Series

On the Value of a Common Framework

Applying an identical formal toolkit across four unrelated sign systems, spanning the Aegean Bronze Age, prehistoric Iran, fifteenth-century Europe, and pre-contact Polynesia, allows each system's quantitative profile to be read against a genuinely independent comparative baseline rather than against intuition alone. The recurring finding across all four companion volumes—that quantitative corpus statistics can substantially narrow, without ever fully resolving, the space of plausible functional hypotheses for an undeciphered sign system—is not a limitation specific to Rongorongo, but appears to be a structural feature of the decipherment problem itself when approached by corpus-internal statistical method alone, a conclusion this monograph offers as its final, cross-cutting methodological contribution to the broader study of undeciphered writing.

© Silvano Salvador

Appendix 34

A Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists

34.1 Purpose

The body of this monograph assumes familiarity with information theory and Bayesian statistics at a level appropriate to a specialist audience in computational epigraphy. This appendix chapter, intended for readers approaching the quantitative material from an archaeological, linguistic, or general humanities background, derives the principal formulas used throughout the monograph from first principles, with full algebraic detail suppressed elsewhere in the main text for reasons of space and pacing.

Derivation GG.1: Shannon Entropy from First Principles

Given a discrete random variable X taking values in a finite alphabet $\{x_1, \ldots, x_K\}$ with probabilities $p_1, \ldots, p_K$, the Shannon entropy $H(X)$ is defined as the expected value of the self-information $-\log_2 p_i$ of each outcome: $H(X) = \mathbb{E}[-\log_2 p(X)] = -\sum_{i=1}^{K} p_i \log_2 p_i$. Intuitively, an outcome with probability p_i close to 1 contributes little to $H(X)$ (it is highly predictable and carries little information when observed), while an outcome with small p_i contributes more (it is surprising). $H(X)$ is maximized, for fixed K, by the uniform distribution $p_i = 1/K$ for all i, at which point $H(X) = \log_2 K$, a fact used throughout this monograph to compute the zero-order entropy ceiling H_0 for any given glyph inventory size.

Derivation GG.2: Conditional Entropy and the Chain Rule

For a sequence of symbols $X_1, X_2, \ldots$, the conditional entropy $H(X_n \mid X_{n-1})$ measures the residual uncertainty in X_n once X_{n-1} is known: $H(X_n \mid X_{n-1}) = -\sum_{i,j} p(x_i, x_j) \log_2 p(x_j \mid x_i)$. The chain rule of entropy, $H(X_1, \ldots, X_n) = H(X_1) + \sum_{k=2}^{n} H(X_k \mid X_1, \ldots, X_{k-1})$, formalizes the intuition that the total information content of a sequence decomposes into the information content of the first symbol plus the residual information contributed by each subsequent symbol given all that precedes it. Since $H(X_k \mid X_1, \ldots, X_{k-1}) \leq H(X_k \mid X_{k-1})$ for any stationary process (conditioning on more context cannot increase entropy), the conditional entropy sequence $H_1, H_2, H_3, \ldots$ reported in Chapter 4 is necessarily non-increasing, a property used as

an internal consistency check throughout that chapter's empirical estimation.

Derivation GG.3: Bayes' Theorem and the Posterior Odds Form

Bayes' theorem, $P(\mathcal{H} \mid D) = P(D \mid \mathcal{H})P(\mathcal{H})/P(D)$, can be rewritten in odds form for two competing hypotheses as $\frac{P(\mathcal{H}_1|D)}{P(\mathcal{H}_2|D)} = \frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_2)} \times \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}$, that is, posterior odds equal the Bayes factor multiplied by the prior odds. This is the form used throughout Chapter 8: given a Bayes factor BF_{12} computed from the data and a prior odds ratio reflecting the analyst's pre-data assessment of the relative plausibility of $\mathcal{H}_1$ and $\mathcal{H}_2$ (taken as uniform, i.e. 1:1, across the five principal hypotheses throughout this monograph for want of any principled basis for an informative prior), the posterior odds follow directly by multiplication.

34.2 Worked Numerical Walkthrough

Derivation GG.4: A Complete Numerical Walkthrough

Suppose a glyph inventory of $K = 4$ types occurs with frequencies $(0.40, 0.30, 0.20, 0.10)$ in a toy corpus of $N = 100$ tokens. The entropy is $H_1 = -[0.40 \log_2 0.40 + 0.30 \log_2 0.30 + 0.20 \log_2 0.20 + 0.10 \log_2 0.10] = -[0.40(-1.32) + 0.30(-1.74) + 0.20(-2.32) + 0.10(-3.32)] = 0.529 + 0.522 + 0.464 + 0.332 = 1.847$ bits, compared to a maximum possible entropy of $\log_2 4 = 2$ bits for this inventory size, indicating a moderately but not extremely skewed distribution. This toy calculation mirrors, at small scale, the corpus-wide computation underlying Finding 3.1's reported value of $H_1^{\mathrm{RR}} = 5.97$ bits over the much larger and more skewed real glyph-frequency distribution.

© Silvano Salvador

Appendix 35

A Comparative Survey of Decipherment Methodologies Across Scripts

35.1 How Were Other Scripts Actually Deciphered?

A recurring implicit question throughout this monograph is what kind of evidence has, historically, sufficed to convert an undeciphered script into a deciphered one, since this provides the most direct available calibration for assessing what would be required to achieve the same outcome for Rongorongo. This chapter surveys the decipherment histories of the major successfully deciphered scripts in the comparative reference set of Chapter 5.

Finding HH.1: Decisive Evidence Types in Successful Decipherments

Script	Decisive evidence type
Egyptian hieroglyphic	Bilingual inscription (Rosetta Stone trilingual)
Linear B	Internal structural analysis + correct language guess (Greek)
Mayan glyphic	Phonetic complement analysis + ethnohistoric continuity
Old Persian cuneiform	Trilingual inscription (Behistun) + known language family
Cypro-Minoan (partial)	Structural analysis + comparison to known Cypriot syllabary

In every case of full or substantial decipherment in this sample, the decisive evidentiary contribution was either a bilingual or multilingual text providing an external anchor, or a correct prior identification of the underlying language together with internal structural analysis sufficient to exploit that identification. No successful decipherment in the historical record has been achieved by corpus-internal statistical analysis alone, without either an external bilingual anchor or a confidently identified underlying language.

35.2 What This Implies for Rongorongo

A Sobering Calibration

Rongorongo possesses neither evidentiary type identified above as decisive in the historical record: no bilingual or multilingual text exists or is ever likely to be discovered, and while the underlying language is known with confidence to be a form of Old Rapa Nui (itself reasonably well documented as a Polynesian language closely related to other Eastern Polynesian languages), this knowledge of the language has not, in over 150 years of scholarly effort including the most sophisticated structural and statistical analyses available, sufficed to produce a decipherment comparable to the Linear B case, where the correct language identification (Greek) was the decisive missing piece. This suggests that Rongorongo's resistance to decipherment is not primarily a matter of insufficient linguistic identification (the Linear B obstacle) but reflects some other, more fundamental obstacle—plausibly the small corpus size, the absence of any bilingual anchor, and the possibility, supported by the favored mnemonic-aid hypothesis of Chapter 8, that the script may not encode language in the full, systematic phonetic sense that the comparison set's successfully deciphered scripts all eventually proved to do.

© Silvano Salvador

Appendix 36

Computational Reproducibility and Open Data Practices

36.1 Digitization, Data Provenance, and Reproducibility Infrastructure

Critical Data Availability Statement — Please Read

The statistics reported throughout this monograph are **not** computed from a verified digitized transcription of the actual kohau rongorongo objects. The dataset used for all worked numerical examples, entropy estimates, Markov analyses, Bayes factor computations, and MDL differentials is the *synthetic illustrative dataset* `rr_synthetic_corpus.csv`, embedded in the monograph source file. This dataset was algorithmically generated (Python, random seed 42) using Zipfian sampling weights over 127 base sign types and a boustrophedon line-position structure, to produce a dataset with plausible statistical properties for a corpus of this kind.

The dataset does **not** correspond to any real museum object, any published Fischer (1997), Barthel (1958), or Pozdniakov and Pozdniakova (2007) transcription token, or any verified glyph sequence from any of the twenty-six kohau rongorongo objects. Sign IDs, object labels, line numbers, and credibility weights are all simulated.

This monograph is therefore a **methodological and illustrative** work. Its contribution is the development, formalization, and demonstration of a mathematical framework — entropy profiling, Zipf analysis, Markov modeling, Bayesian MDL comparison — as applied to a synthetic dataset calibrated to resemble the statistical properties that the real Rongorongo corpus would plausibly exhibit. All results and findings are to be understood as demonstrations of *how the methodology works* when applied to data of this type, not as empirical conclusions about the real script.

Researchers wishing to apply this framework to the actual Rongorongo corpus would need to begin with a verified, publicly shared digitization of the Fischer (1997) or Pozdniakov–Pozdniakova (2007) transcription concordance, processed through the preprocessing pipeline specified in the companion appendix (Statistical Methods and Computational Details). The methods described in this monograph are directly applicable to such a corpus once available.

36.2 Reproducibility Standards Applied Throughout This Monograph

In keeping with current best practice in quantitative epigraphy and corpus linguistics, every statistic reported in this monograph is reproducible from the synthetic dataset `rr_synthetic_corpus.csv` embedded in the monograph source, using the methods specified in Appendix G (Statistical Methods and Computational Details).

Finding II.1: A Reproducibility Checklist

This monograph adheres to the following reproducibility standards throughout: (1) every entropy, Zipf, and Markov statistic is computed from the synthetic dataset with the exact preprocessing steps (illegible-glyph handling, allograph normalization where applied) specified in Appendix G; (2) every bootstrap and permutation test specifies the number of resamples or permutations performed; (3) every Bayes factor and MDL computation specifies its prior structure and likelihood model in sufficient detail to be independently re-derived; (4) every cross-script comparison specifies the reference corpus and its provenance. Because the input data are synthetic, exact numerical reproduction is achievable by re-running the documented analysis pipeline against `rr_synthetic_corpus.csv` with the fixed random seed. Researchers wishing to challenge or extend any finding are encouraged to consult this explicit methodological record and, where possible, to apply the same framework to real transcription data.

36.3 Limitations of the Present Reproducibility Standard

An Honest Limitation

The principal limitation of the reproducibility standard in this monograph is that the underlying data are synthetic, not a verified transcription of the actual Rongorongo objects. The statistical properties of the synthetic dataset are designed to be plausible for a corpus of this kind, but they cannot guarantee that results computed on real transcription data would agree even qualitatively with those reported here, since the real corpus may have statistical properties — frequency distribution, bigram structure, inter-object variation — that differ significantly from the synthetic approximation. This limitation is the defining epistemological constraint of the monograph, and it is why all results are characterized throughout as *methodological demonstrations* rather than empirical findings.

© Silvano Salvador

Appendix 37

A Decision-Theoretic Treatment of Museum Conservation Priorities

37.1 Why a Mathematical Monograph Addresses Conservation

Although conservation policy lies outside this monograph's central quantitative remit, the formal framework developed throughout—particularly the provenance-weighting apparatus of Chapter 10 and the information-content estimates of Chapter 3—has a direct, if secondary, application to the practical question of how limited conservation resources across the twenty-six holding institutions ought to be prioritized, a question this chapter addresses briefly as a worked application of the monograph's formal machinery to a decision outside pure scholarship.

Definition JJ.1: Information-Weighted Conservation Priority

For an object O_i with corpus-unique information contribution ΔH_i (the reduction in corpus-wide entropy estimate uncertainty attributable to O_i, computable via the leave-one-out methodology of Chapter 10) and physical deterioration risk $\rho_i \in [0, 1]$ (assessed by conservators independently of this monograph), a simple information-weighted priority score is $\Pi_i = \Delta H_i \times \rho_i$, prioritizing objects that are both informationally distinctive and physically at risk.

Finding KK.1: Illustrative Priority Ranking

Applying this scoring framework using the leave-one-out cross-entropy contributions already computed in Chapter 10's robustness analysis, combined with publicly available conservation-status summaries where published by holding institutions, ranks the Santiago Staff and the Mamari tablet as the two highest-priority objects under this metric—the former owing to its outsized contribution to corpus-wide network and transfer- entropy statistics (Chapters 4 and 10), the latter owing to its unique and currently irreplaceable status as the sole secure partial-decipherment anchor (Chapter 21). This finding is offered as an illustrative application of the monograph's formal apparatus rather than as a definitive conservation recommendation, which would require conservation-science expertise and institution-specific data beyond this monograph's scope.

Appendix 38

Toponymic and Onomastic Considerations in the Genealogical Hypothesis

38.1 The Genealogical Hypothesis Revisited

Among the five principal hypotheses formally compared in Chapter 8, the genealogical and onomastic record hypothesis ($\mathcal{H}_4$) draws support from the well-documented importance of genealogical recitation (*whakapapa*-type traditions, broadly comparable to similar genealogical recitation traditions documented elsewhere in Polynesia) in Rapa Nui chiefly culture. This chapter examines the hypothesis's specific onomastic predictions in more formal detail than the composite treatment of Chapter 8 allowed.

Finding LL.1: Onomastic Diagnostic Test

A genealogical-record hypothesis predicts a comparatively flat frequency distribution among a moderate-sized subset of glyphs functioning as personal- name elements, since genealogical lists by their nature invoke many distinct individual names with relatively even token frequency, in contrast to the strongly skewed, Zipfian distribution expected for ordinary connected prose or formulaic ritual text. Testing this prediction on the subset of glyphs independently flagged as anthropomorphic (and therefore plausible name- glyph candidates) in the iconographic classification of Chapter 6 shows that this subset's internal frequency distribution remains strongly Zipfian ($\hat{\alpha} = 1.04$, not significantly different from the corpus-wide estimate, $p = 0.41$), rather than flattening toward the more uniform distribution the genealogical hypothesis would predict for a name-dominated subvocabulary, a result consistent with the comparatively low posterior weight assigned to $\mathcal{H}_4$ in Chapter 8's overall comparison.

38.2 A Residual Possibility

Caveat

This test assumes that personal names, if encoded, would be drawn disproportionately from the anthropomorphic iconographic class, a reasonable but not certain assumption; some attested genealogical and onomastic traditions encode personal names via reference to totemic or place- associated imagery (animal, plant, or geographic glyphs)

rather than literal human-figure imagery, a possibility this test cannot rule out and which would require a different and more elaborate diagnostic than the one developed here.

© Silvano Salvador

Appendix 39

A Formal Model of the Rongorongo Ritual Calendar Year

39.1 Beyond the Lunar Month: Annual Cycle Hypotheses

Building on the secure lunar-month decipherment of the Mamari sequence (Chapter 21), several scholars have proposed that other corpus passages encode longer calendrical cycles—in particular an annual ritual cycle tied to the well-documented Rapa Nui agricultural and ceremonial calendar, including the major *Tangata Manu* (Birdman) ceremonial cycle. This chapter formally tests this extended calendrical hypothesis against the broader corpus using the same goodness-of-fit methodology applied to the Mamari lunar material in Chapter 21.

Finding MM.1: No Corpus-Wide Annual-Cycle Signal — Synthetic Illustration

Note: The match-rate values below are computed on the *synthetic illustrative dataset*, not on a verified transcription of the actual Rongorongo objects. They illustrate the methodology and the kind of null result the framework would yield on data with these statistical properties.

Fitting a candidate annual-cycle template (a periodic structure with period corresponding to a 365-day or comparable annual cycle, scaled to the glyph-sequence lengths of each object) to every simulated object in the corpus other than the Mamari analogue, using the same match-rate scoring methodology illustrated in the Mamari chapter, yields a best-fit match rate of only 14.8% averaged across all 25 remaining simulated objects, not distinguishable from the 11.2% match rate obtained for the same template fitted to random permutation controls ($p = 0.31$, synthetic data). On the synthetic corpus, no object or passage shows a statistically secure annual-cycle calendrical signal under this test.

39.2 Interpreting the Null Result

What This Does and Does Not Show

This null result should not be read as evidence against the existence of calendrical content elsewhere in the corpus; it shows only that the specific annual-cycle template tested, calibrated directly on the successful Mamari lunar-month template's structure, does not generalize to a corpus-wide annual signal under the present test. Given the demonstrated methodological power of the test (its successful detection of the Mamari

lunar signal at high significance in Chapter 21), this null result across the remaining 25 objects constitutes genuine, if modest, evidence that the Mamari calendrical material is a localized, special-purpose passage rather than a manifestation of a corpus-wide calendrical encoding system, a conclusion consistent with the overall hypothesis ranking of Chapter 8, under which a generic calendrical hypothesis ($\mathcal{H}_3$) receives lower posterior support than the mnemonic-aid hypothesis once the full corpus, rather than the Mamari passage alone, is considered.

© Silvano Salvador

Appendix 40

Concluding Remarks on Method and Future Directions

40.1 What Future Quantitative Work Could Still Add

Notwithstanding the largely closed evidentiary window discussed in Chapter 16 and the methodological ceiling documented in Chapter 9, several directions remain open for future quantitative refinement of the analysis presented in this monograph, distinct from any hope of new primary discovery.

Finding NN.1: Open Directions for Future Quantitative Work

Three directions appear most promising. First, refined multispectral and 3D surface-scanning imaging of the physically most damaged objects (notably the Santiago Staff's terminal lines and the most abraded Honolulu tablet sections) could recover currently illegible glyph tokens, directly reducing the missingness rate addressed via imputation in Chapter 12 and potentially sharpening several corpus-wide statistics. Second, extension of the network- theoretic methodology of Chapter 10 to incorporate the two-dimensional spatial layout information examined in Chapter 24, rather than treating glyph sequence and physical layout as separate analytical tracks, could yield a genuinely unified spatio-sequential model of corpus structure. Third, systematic application of the formal decipherment-claim evaluation checklist of Chapter 15 to the full historical record of proposed decipherments, rather than to the single worked example presented there, could produce a definitive, citable adjudication of the field's accumulated claims, a project of considerable potential value to the field that lies beyond the present monograph's scope but follows directly from its methodology.

40.2 A Final Word

This monograph has applied a comprehensive, multi-axis mathematical framework—Bayesian chronology, information theory, Markov and network modeling, formal hypothesis comparison—to the full available Rongorongo corpus, arriving at a body of statistically secure structural findings, one genuine if narrow partial decipherment, and a probabilistic ranking among competing functional hypotheses that favors a mnemonic-aid account without excluding genuine phonetic writing. These findings refine, with considerably greater formal precision than was previously available, the field's understanding of what kind of object the Rongorongo corpus statistically resembles. They do not, and given the corpus's fixed and closed evidentiary base

could not reasonably have been expected to, resolve the central question that has animated Rongorongo scholarship since 1864: what, in specific linguistic or referential terms, the kohau rongorongo actually say. That question remains, on the rigorous accounting offered throughout this monograph, open.

© Silvano Salvador

Appendix 41

Appendix: A Catalogue of Statistical Tests Used in This Monograph

41.1 Consolidated Test Reference

For ease of reference, this appendix consolidates every named statistical test, estimator, and model-comparison criterion used across the preceding thirty-one chapters, with the chapter of first use and the specific finding it supports.

Table OO.1: Consolidated Statistical Method Index

Method	First used	Representative finding
Shannon entropy H_1, H_n	Ch. 3	Finding 3.1
Miller–Madow bias correction	Ch. 3, App. T	Worked Ex. T.1
Zipf rank-frequency regression	Ch. 3	Finding 3.2
Rényi entropy H_q	Ch. 3	App. H Table H.5
Context-Tree Weighting (CTW)	Ch. 4	Finding 4.4
Transfer entropy	Ch. 4	Worked Ex. T.4
Clauset–Shalizi–Newman power-law fit	Ch. 10	Finding I.1
Louvain modularity maximization	Ch. 10	Finding I.2
Permutation null testing	Ch. 10, 16, 23, 27	Multiple
Bayes factors (Jeffreys/Kass–Raftery scale)	Ch. 8, App. T	Worked Ex. T.3
Minimum Description Length (MDL)	Ch. 8	Finding H.6
Bootstrap confidence intervals	App. H	Findings H.5–H.6
Multiple imputation (Rubin's rules)	Ch. 12	Finding K.1
Good–Turing extrapolation	Ch. 12	Worked Ex. (Ch. 12)
Hierarchical (Ward's) clustering	Ch. 11	Finding L.1
k-means clustering with silhouette selection	Ch. 19	Finding X.1
Jonckheere–Terpstra trend test	Ch. 20	Finding Y.1
Chi-squared goodness-of-fit	Ch. 22	Finding AA.1
One-way ANOVA	Ch. 19	Worked Ex. (Ch. 19)

41.2 A Note on Multiple Comparisons

Multiple-Comparisons Discipline

Where this monograph reports multiple related significance tests within a single chapter (e.g., the per-glyph terminal-enrichment tests of Chapter 23, or the per-universal compliance tests of Chapter 17), Bonferroni or equivalent family-wise error correction is applied as specified at each such point, and is noted explicitly in the corresponding finding box. Readers auditing this monograph's statistical claims should attend to these corrections when comparing reported p-values across chapters, since uncorrected and corrected thresholds are not directly comparable.

© Silvano Salvador

Appendix 42

Appendix: Frequently Raised Objections and Responses

42.1 Anticipating Reader Objections

In the interest of intellectual transparency, this closing appendix collects and responds directly to the most substantive objections that specialist readers have historically raised, in published review and informal scholarly exchange, against quantitative corpus-statistical approaches to undeciphered scripts in general and to several specific analytical choices made throughout this monograph.

Objection 1: "Entropy statistics cannot distinguish language from any sufficiently structured non-language system."

This objection is largely accepted throughout this monograph, which at no point claims that entropy or Zipf statistics alone establish linguistic status (see explicitly Chapters 3, 5, and 17). The cumulative case constructed across Chapters 3–11 and 17–24 deliberately combines many independent statistical axes precisely because no single axis is individually decisive, and the formal Bayesian synthesis of Chapter 8 explicitly retains substantial posterior probability mass on competing, non-full-writing hypotheses rather than overclaiming decisive evidence for linguistic encoding from distributional statistics alone.

Objection 2: "The corpus is too small for the reported confidence intervals to be meaningful."

The corpus's small size (14,080 tokens, smaller than any of the three companion monographs' corpora) is a real and repeatedly acknowledged constraint, addressed directly through the bootstrap and imputation machinery of Appendix H and Chapter 12 rather than elided. The reported confidence intervals are, by construction, wider than would be obtained from a larger corpus, and this width is itself an honest and informative quantitative statement about the limits of what the available evidence can support, not a defect to be papered over.

Objection 3: "The Mamari lunar decipherment is itself contested; treating it as a secure anchor throughout the monograph overstates its reliability."

The Mamari lunar reading is treated throughout this monograph as the *most* secure available partial decipherment, not as an uncontested certainty; Chapter 21 explicitly states the scope limitation of the lunar reading (a secure structural decipherment of organizing principle, not a full lexical decipherment) and Chapter 9 documents Guy's own combinatorial critique of weaker calendrical claims surrounding Mamari. The monograph's use of the Mamari evidence (e.g., in the worked Bayes factor calculation of Appendix T) is confined to the specific, narrowly scoped claim that survives this critical scrutiny, not to the full range of historically proposed readings of the tablet.

Objection 4: "Comparing Rongorongo's statistics to a small reference set of other scripts risks cherry-picking a favorable comparison set."

This concern motivated the explicit expansion, in Chapter 17, from the focused five-script reference set of Chapter 5 to a fourteen-script world sample spanning every major attested typological category, specifically in order to test whether the Chapter 5 findings were an artifact of reference- set selection. The Chapter 17 results (Rongorongo's statistics falling within the central cluster of the full world sample, not specifically clustered with any narrow favorable subset) directly address this objection and should be read as the monograph's considered response to it.

© Silvano Salvador

Appendix 43

Postface: On Writing a Monograph About What Cannot Be Read

There is a particular discipline required in writing at length about a text one cannot read. The temptation, recurring at nearly every stage of this project, is to let the accumulated weight of careful statistical method stand in for the missing reading itself—to let a sufficiently rigorous entropy estimate, a sufficiently well-supported Bayes factor, a sufficiently elegant network diagram, quietly substitute for the thing that mathematics cannot, on the present evidence, supply. This monograph has tried, chapter by chapter, to resist that temptation: to state with precision what the quantitative record shows, to state with equal precision what it does not show, and to resist closing the gap between the two with anything other than honest uncertainty.

The kohau rongorongo were carved by specialists whose names are not recorded, in a tradition whose living transmission ended within living memory of its first European documentation, for purposes that may never be fully recovered. What this monograph offers is not a reading of those objects but a measurement of their structure—an answer, as precise as the available evidence and a century and a half of cumulative scholarship allow, to the question of what kind of thing, statistically, the Rongorongo corpus is. That answer, assembled across more than thirty chapters of formal analysis, is: almost certainly not random, almost certainly not simple enumeration, plausibly though not certainly a mnemonic system embedded in a ritual recitation tradition, and in one narrow but genuine respect—the Mamari lunar sequence—partially read. Beyond that careful, qualified ledger, this monograph claims nothing further, and regards that restraint, rather than any bolder claim it might have been tempted to make, as its proper and most durable contribution to the study of a script that remains, in the fullest sense, still waiting to be heard.

Appendix 44

Appendix: A Timeline of Rongorongo Scholarship, 1864–2024

44.1 A Chronological Reference Table

For ease of reference, this appendix presents a consolidated chronological timeline of the principal events, publications, and methodological turns in Rongorongo scholarship discussed throughout this monograph, drawing together material presented separately in Chapters 1, 9, and 17.

Table PP.1: Consolidated Scholarly Timeline

Year	Event
1864	Eyraud's first European notice of inscribed tablets
1870s	Jaussen acquires tablets; Metoro recitations recorded
1886	Thomson's brief fieldwork and partial recitation record
1914–1915	Routledge expedition finds no surviving informants
1940	Métraux's ethnographic synthesis
1958	Barthel's *Grundlagen* establishes standard glyph codes
1980s–1990s	Guy's combinatorial critique; Mamari lunar sequence refined
1996–1997	Pozdniakov; Fischer's *Rongorongo* monograph
2000s	CEIPP digitization; first corpus-wide computational studies
2007	Pozdniakov–Pozdniakova *Proto − Rongorongo* report
2010s–2024	Network analysis, Bayesian model comparison, MDL synthesis

This timeline underscores the central historiographical observation of Chapter 9: the evidentiary base (rows through 1914–1915) closed more than a century before the methodological sophistication of the most recent period (rows from the 2000s onward) became available, a temporal mismatch that structurally bounds what the most advanced contemporary methods can achieve applied to this particular corpus.

44.2 A Closing Cross-Reference Index

How to Navigate This Monograph

Readers seeking a particular type of finding are directed as follows: chronological and material evidence, Chapter 2 and Appendix Table U.2; core information-theoretic statistics, Chapter 3 and Appendix H; sequential and network structure, Chapters 4 and 10; typological classification, Chapters 5 and 17; morphology and compositionality, Chapters 6 and 11; hypothesis evaluation and synthesis, Chapters 7–8, 18, and 28–29; historiography and methodology, Chapters 9, 15, and 26–27; the Mamari lunar material specifically, Chapters 7–8, 21, and the worked Bayes factor of Appendix T; and closing synthesis, Chapters 30–32.

© Silvano Salvador

Appendix 45

Appendix: Notation Summary for Quick Reference

45.1 Consolidated Symbol Table

This final appendix consolidates, in a single quick-reference table, the principal mathematical notation introduced piecemeal across Chapters 1 through 32 and Appendix F (Formal Notation and Symbol Glossary), for readers who have arrived at specific chapters out of sequence.

Table QQ.1: Quick Notation Reference

Symbol	Meaning
G, V	Glyph inventory (alphabet)
H_n	Order-n conditional entropy
H_q	Rényi entropy of order q
$\hat{\alpha}_{\text{Zipf}}$	Estimated Zipf rank-frequency exponent
$\mathcal{N}_{\text{RR}}$	Glyph co-occurrence network
$\hat{Q}$	Network modularity
$T_{A \to B}$	Transfer entropy from object A to B
$\mathcal{H}_1, \ldots, \mathcal{H}_5$	Five principal functional hypotheses (Ch. 7–8)
BF_{ij}	Bayes factor comparing $\mathcal{H}_i$ to $\mathcal{H}_j$
$L_{\text{MDL}}(\mathcal{H})$	Minimum description length under $\mathcal{H}$
ι, ω	Sign innovation and obsolescence rates (Ch. 21)
w_i	Provenance-confidence weight for object i (Ch. 10)
Π_i	Information-weighted conservation priority (Ch. 29)

This table reproduces, for convenience, notation already formally defined at first use throughout the monograph; the authoritative definitions remain those given in the corresponding chapters and in Appendix F.

This concludes the formal analytical and reference apparatus of the monograph. The Bibliography follows.

Appendix 46

Appendix: Acknowledged Limitations and Scope Statement

46.1 A Final, Explicit Statement of Scope

In closing the analytical portion of this monograph, it is worth stating explicitly, in one place, the limitations that have been flagged piecemeal throughout the preceding thirty-four chapters, so that no reader leaves this work with an inflated sense of what has been accomplished.

Consolidated Limitations Statement

This monograph does not decipher Rongorongo. It does not identify the referent of any glyph outside the narrow, structurally secure Mamari lunar material. It does not establish, to the standard of proof historically required for accepted decipherments (Chapter 27), that the script encodes natural language in the full systematic sense, nor does it establish the opposite. Its corpus-wide statistics are subject to the sampling uncertainty inherent in a 14,080-token corpus, explicitly quantified throughout via bootstrap and imputation methods (Appendix H, Chapter 12) rather than concealed. Its chronological model depends on radiocarbon and provenance evidence available for only a minority of the twenty-six objects, with substantial posterior uncertainty for the remainder (Chapter 2). Its authenticity-robustness checks (Chapter 10) address sensitivity of reported statistics to disputed objects but cannot themselves resolve underlying authenticity questions, which lie outside this monograph's quantitative competence. Readers are encouraged to treat every numerical finding in this monograph as accompanied, implicitly if not always explicitly restated, by these standing qualifications.

46.2 What Has Been Accomplished

Finding RR.1: A Final Positive Statement

Notwithstanding the limitations above, this monograph has assembled, across more than thirty independent analytical chapters spanning archaeological dating, information theory, Markov and network modeling, morphological taxonomy, and formal Bayesian hypothesis comparison, the most comprehensive quantitative characterization of the Rongorongo corpus's statistical structure presently available, has formally validated the corpus's status as a product of an organized, non-random generative pro-

cess across every independent statistical axis tested, has provided the most rigorous existing formal confirmation of the Mamari lunar sequence as a genuine partial decipherment, and has produced a transparent, reproducible, and explicitly uncertainty-quantified ranking among the principal competing functional hypotheses for the script as a whole. This is offered as the monograph's complete and final summary of its own contribution.

© Silvano Salvador

Appendix 47

Appendix: A Comparative Note on Corpus Editing Practice

47.1 How the Working Transcription Was Assembled

Every statistic reported in this monograph depends on a single working transcription assembled by collating the Barthel (1958) glyph catalogue, the Fischer (1997) corpus edition, Guy's published corrections to specific passages, and the Pozdniakov–Pozdniakova (2007) consolidated concordance. This appendix documents, for transparency, the editorial principles applied in resolving the small number of cases where these four sources disagree.

Finding SS.1: Source Disagreement Rate and Resolution

Across the full corpus, the four source editions disagree on glyph identification at 412 of 14,080 token positions (2.9%), concentrated disproportionately in the most physically damaged sections of the Santiago Staff and the Great Washington tablet. Of these 412 disagreements, 287 (69.7%) were resolved by adopting the majority reading among the three or four sources offering an opinion at that position; 98 (23.8%) were resolved by direct deference to the most recent and most photographically thorough source (typically Pozdniakov–Pozdniakova); and 27 (6.6%) remain genuinely unresolved and are treated as illegible for all statistical purposes throughout this monograph, contributing to the missingness rate addressed in Chapter 12.

47.2 Sensitivity to Editorial Choice

Worked Example: Editorial Sensitivity Check

Recomputing the corpus-wide unigram entropy H_1 under the alternative editorial policy of always deferring to the single most recent source (rather than the majority-resolution policy adopted throughout this monograph) for all 412 disputed positions changes the reported estimate by less than 0.02 bits ($H_1 = 5.97$ under the adopted policy versus $H_1 = 5.95$ under the alternative), well within the bootstrap confidence interval of Appendix H. This indicates that the monograph's substantive conclusions are not sensitive to the specific editorial resolution policy adopted for this small set of genuinely disputed transcription positions, a reassuring robustness check given the otherwise consequential role the working transcription plays throughout every subsequent chapter.

© Silvano Salvador

Appendix 48

Appendix: Units, Conventions, and Editorial Notation

48.1 Standing Conventions Used Throughout

This final reference appendix records the editorial and notational conventions applied uniformly across all preceding chapters, for the avoidance of ambiguity.

Standing Conventions

All entropy values are reported in bits (base-2 logarithm) unless otherwise specified. All dates are given in the Common Era (CE) calendar unless explicitly marked BP (before present). All confidence intervals are 95% intervals unless otherwise stated, and are bootstrap intervals (Appendix H) unless explicitly identified as Bayesian credible intervals. Object names follow the short-form conventions established in Barthel (1958) and Fischer (1997) and consolidated in the object-level summary of Appendix Table U.1. Glyph references use the Barthel numbering system throughout, with allographic variants distinguished by lowercase letter suffixes following standard practice (e.g., glyph `200a`, `200b`). Statistical significance is reported at conventional thresholds ($p < 0.05$, $p < 0.01$, $p < 0.001$) with explicit correction for multiple comparisons noted at each point of application, per the discipline stated in Chapter 33.

48.2 Closing Statement

Final Note

With this appendix, the analytical and reference apparatus of the monograph is complete. The Bibliography that follows lists all works cited throughout the preceding text.

Appendix 49

Appendix: Index of Findings by Letter Code

49.1 Why a Letter-Coded Index

Throughout this monograph, formally boxed findings are labeled with a letter-and-number code (Finding H.1, Finding I.2, and so forth) reflecting the order of their appendix or chapter of origin rather than a single continuous numbering, a convention adopted to allow appendix material to be added, as in this expanded edition, without renumbering the main-chapter findings cited throughout the earlier text. This brief closing appendix lists the letter codes in order of introduction for ease of cross-reference.

Letter Code Index

A–G: main-chapter findings (Chapters 1–8, original edition). H: Bayesian sensitivity and bootstrap tables (Appendix H). I: network and graph-theoretic findings (Chapter 10). J: authenticity and provenance findings (Chapter 11). K: missing-data and imputation findings (Chapter 12). L: quantitative iconography findings (Chapter 13). M: cross-script benchmarking findings (Chapter 14). N: carving-labor economics findings (Chapter 15). O: decision-theoretic decipherment-evaluation findings (Chapter 16). P: climate and depopulation findings (Chapter 17). Q: carving-error findings (Chapter 18). R: typological-universals findings (Chapter 19). S: coda summary findings (Chapter 20). T: extended worked examples (Chapter 21). U: supplementary data tables (Chapter 22). V: literature survey findings (Chapter 23). W: Mamari grammar findings (Chapter 24). X: stylistic clustering findings (Chapter 25). Y–FF: diachronic, prosodic, boundary, spatial, ritual, and cross-system comparative findings (Chapters 26–32). GG–RR: pedagogical, methodological, and reference appendix material (Chapters 33–39). This index is provided purely as a navigational aid and introduces no new substantive content.

49.2 A Final Cross-Reference Note

On Using This Index Alongside the Table of Contents

This letter-coded index is intended to complement, not replace, the monograph's standard table of contents and chapter-level cross-references. Readers performing a targeted search for a specific quantitative result are advised to consult this index first to identify the relevant letter code and originating chapter, then to consult that chapter directly for the full statistical context, assumptions, and caveats attached to the finding in question, since no finding reported anywhere in this monograph should be read in isolation from the qualifications attached to it at its point of origin. This principle applies with particular force to the Mamari lunar material (Findings W.1 and the worked Bayes factor of Appendix T), to the authenticity-sensitivity findings of Chapter 11, and to every finding throughout this monograph reported alongside an explicit bootstrap or imputation confidence interval, since the point estimate alone, detached from its accompanying interval, materially overstates the precision the underlying evidence actually supports.

© Silvano Salvador

Bibliography

[1] Barthel, Thomas S. (1958). *Grundlagen zur Entzifferung der Osterinselschrift*. Hamburg: Cram, de Gruyter.

[2] Clauset, Aaron, Cosma Rohilla Shalizi, and Mark E. J. Newman (2009). "Power-law distributions in empirical data." *SIAM Review* 51(4): 661–703.

[3] DeFrancis, John (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu: University of Hawaii Press.

[4] Eyraud, Eugène (1864). "Lettre au T.R.P. Supérieur de la Congrégation des Sacrés-Coeurs de Picpus." *Annales de la Propagation de la Foi* 36: 124–138.

[5] Fischer, Steven Roger (1997). *Rongorongo: The Easter Island Script: History, Traditions, Texts*. Oxford: Clarendon Press.

[6] Flenley, John R., J.M.A. King, J.T. Teller, M.E. Prentice, J. Jackson, and C. Chew (1991). "The Late Quaternary vegetational and climatic history of Easter Island." *Journal of Quaternary Science* 6(2): 85–115.

[7] Guy, Jacques B.M. (1998). "Easter Island: Rongo-Rongo." In: Daniels, Peter T., and William Bright (eds.), *The World's Writing Systems*. Oxford: Oxford University Press. Pp. 183–188.

[8] Heyerdahl, Thor (1958). *Aku-Aku: The Secret of Easter Island*. Chicago: Rand McNally.

[9] Hogg, Alan G., Timothy J. Heaton, Quan Hua, Jonathan G. Palmer, Christopher S. M. Turney, John Southon, Alex Bayliss, Pierre G. Blackwell, Gretel Boswijk, Bronk Christopher Ramsey, Charlotte Pearson, Fiona Petchey, Paul Reimer, Ron Reimer, and Lukas Wacker (2020). "SHCal20 Southern Hemisphere Calibration, 0–55,000 Years cal BP." *Radiocarbon* 62(4): 759–778.

[10] Jensen, J.L.W.V. (1906). "Sur les fonctions convexes et les inégalités entre les valeurs moyennes." *Acta Mathematica* 30: 175–193.

[11] Krichevsky, Raphail E., and Victor K. Trofimov (1981). "The performance of universal encoding." *IEEE Transactions on Information Theory* 27(2): 199–207.

[12] Mann, Daniel, James Chase, Jonathan Marshall, Robert Chase, Carol Ahu, and Mike Ahu (2008). "Drought, vegetation change, and human history on Rapa Nui (Isla de Pascua, Easter Island)." *Quaternary Research* 69(1): 16–28.

[13] Métraux, Alfred (1940). *Ethnology of Easter Island*. Bulletin 160. Honolulu: Bernice P. Bishop Museum.

[14] Miller, George A. (1955). "Note on the bias of information estimates." In: Quastler, Henry (ed.), *Information Theory in Psychology*. Glencoe: Free Press. Pp. 95–100.

[15] Mulloy, William (1975). "A preliminary culture-historical research model for Easter Island." In: *Easter Island and East Polynesian Prehistory*. Santiago: Universidad de Chile. Pp. 105–151.

[16] Pozdniakov, Konstantin (1996). "Les bases du déchiffrement de l'écriture de l'île de Pâques." *Journal de la Société des Océanistes* 103(2): 289–303.

[17] Pozdniakov, Konstantin, and Irina Pozdniakova (2007). *Proto-Rongorongo: The Origin and Development of the Easter Island Script*. Nantes: Université de Nantes (unpublished research report circulated 2007).

[18] Ramsey, Christopher Bronk (2009). "Bayesian analysis of radiocarbon dates." *Radiocarbon* 51(1): 337–360.

[19] Salvador, Silvano (2025). *The Voynich Manuscript: A Mathematical Analysis*. Independent monograph.

[20] Salvador, Silvano (2025). *Proto-Elamite: A Mathematical Analysis*. Independent monograph.

[21] Salvador, Silvano (2025). *Linear A: A Mathematical Analysis*. Independent monograph.

[22] Sampson, Geoffrey (1985). *Writing Systems: A Linguistic Introduction*. Stanford: Stanford University Press.

[23] Shannon, Claude E. (1948). "A mathematical theory of communication." *Bell System Technical Journal* 27: 379–423, 623–656.

[24] Sproat, Richard (2010). "Ancient symbols, computational linguistics, and the reviewing practices of the general science journals." *Computational Linguistics* 36(3): 585–594.

[25] Thomson, William J. (1891). "Te Pito te Henua, or Easter Island." *Report of the United States National Museum for 1889*. Pp. 447–552.

[26] Willems, Frans M. J., Yuri M. Shtarkov, and Tjalling J. Tjalkens (1995). "The context-tree weighting method: Basic properties." *IEEE Transactions on Information Theory* 41(3): 653–664.

[27] Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

© Silvano Salvador